ERIC Document Resume	**See Instruction Sheet**	**DATE OF RESUME** 6/23/65	**DATE RECEIVED**

A. PUBLICATION DATA

1. ERIC NUMBER ED001917

2. ORIGINATOR CONTROL NO(S). N.A.

3. IDENTIFICATION NO. CR F-025

DO NOT WRITE IN SPACE BELOW

4. AUTHOR(S): Russell, David H.

MICROFILM COPY AVAILABLE YES ☒ NO ☐

5. TITLE OF DOCUMENT ABSTRACTED: Research Design and the Teaching of English

COPYRIGHT GRANTED YES ☒ NO ☐

COPYRIGHT RELEASE FILED YES ☒ NO ☐

6. PUBLICATION TITLE: N.A.

TYPE OF REPORT
- PROPOSAL ☐
- IN-PROGRESS ☒
- FINAL ☐
- FOLLOW-UP ☐

7. EDITOR(S):

8. PLACE OF ORIGIN
ORG. National Council of Teachers of English
ST. 508 South 6th Street
CITY, STATE Champaign, Illinois
TELE. EXT.
INST/LOC CODE

9. PUBLISHER OR SOURCE Cooperative Research Program, Office of Education Department of Health, Education, & Welfare, Wash., D.C.

OTHER ERIC REFERENCES:

ERIC SATELLITE CODE

10. VOL.	11. PAGES	12. MO./YR.	13. TOTAL PP.

B. INDEXING DATA

14. SUBJECT AREA

15. TYPE OF ENDEAVOR Conference

16. INDEXING TERMS (6-12)

	literature	teaching techniques		
research design	composition			
English	reading			
language	language arts			

17. ABSTRACT (MAX. 250 WORDS)

Purposes were: to explore ideas for research, in English, that had emerged at previous conferences; to examine methods of research in literacy scholarship and linguistic science; to consult with experts in research design; and to test the hypothesis that experts in literature, language, psychology, and education can work together effectively to improve knowledge about teaching English.

Two measures were proposed by the conference planners. The first measure was to invite representatives of related disciplines to prepare papers for presentation at the general sessions. The presentations concerned research methods of literacy scholarship, research methods of the linguist, problems and possibilities of research in teaching composition, and research designs for teaching reading in elementary schools and literature in secondary schools. The second measure was the use of study groups composed of research consultants, teaching specialists, and literature scholars.

ERIC No.

18. SOURCE OF ABSTRACT

C. FUNDING DATA

19. SUPPORTING ORGANIZATION

20. DIVISION OR DEPT.

21. CONTRACT NO.

PROJECT NO.

22. TOTAL COST OF PROJECT

23. DURATION OR PROJECT

U.S. Department of Health, Education, and Welfare Office of Education

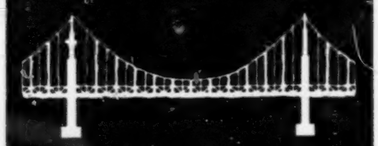
3

ERIC Form 1

F-025

ED001917

Research Design and the Teaching of English



Proceedings of the San Francisco Conference

David H. Russell
Director
Margaret J. Early
Chairman
Edmund Farrell
Assistant
Director

NATIONAL COUNCIL OF TEACHERS OF ENGLISH
508 South Sixth St. • Champaign, Illinois

REPRINTED BY PERMISSION
U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN PRINTED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

ED001917

RESEARCH DESIGN AND THE TEACHING OF ENGLISH

Proceedings of the San Francisco Conference

1963

David H. Russell, Director
Edmund J. Farrell, Assistant Director
University of California

Margaret J. Early, Chairman
Syracuse University

Sponsored by the
University of California, Berkeley

The conference reported herein was supported through the Cooperative Research Program of the Office of Education, United States Department of Health, Education, and Welfare.

NATIONAL COUNCIL OF TEACHERS OF ENGLISH
Champaign, Illinois
1964

Copyright 1964

National Council of Teachers of English

Editorial Committee

Research Design and the Teaching of English

David H. Russell, University of California, Berkeley
Edmund J. Farrell, University of California, Berkeley
Margaret J. Early, Syracuse University
Doris V. Gunderson, United States Office of Education
Richard Braddock, University of Iowa
James R. Squire, *Executive Secretary, NCTE*

Research Committee, NCTE

Margaret J. Early, Chairman, 1959-1963
Doris V. Gunderson, Associate Chairman
Dwight L. Burton, Chairman, 1964-

Richard S. Alm
Mary Austin
Alvina Treut Burrows
John B. Carroll
Edmund J. Farrell
Alexander Frazier
Patrick J Groff
Theodore Harris
J. N. Hook
Henry C. Meckel
Walter T. Petty
William D. Sheldon
J. Harlan Shores
Ingrid Strom
Robert Wright
David H. Russell

NCTE Committee on Publications

James R. Squire, NCTE Executive Secretary, Chairman
Jarvis E. Bush, Wisconsin State University, Oshkosh
Glenn Leggett, University of Washington
Virginia M. Reid, Oakland Public Schools, California
Enid M. Olson, NCTE Director of Publications

ACKNOWLEDGMENTS

This conference was planned by the Committee on Research of the National Council of Teachers of English (Margaret J. Early, Chairman) in cooperation with the Director and the Assistant Director of the conference and the Executive Secretary of NCTE. It was held at the Sheraton-Palace Hotel in San Francisco on November 25 to 27, 1963. The conference was made possible by the combined efforts of:

RICHARD BRADDOCK, Associate Professor of English, University of Iowa; co-author, *Research in Written Composition.*

JOHN B. CARROLL, Professor of Education, Harvard Graduate School of Education.

THEODORE CLYMER, Professor of Education, University of Minnesota; president, International Reading Association (1964-1965) and the National Conference on Research in English (1964-1965).

PAUL B. DIEDERICH, Director of Research in English, Educational Testing Service; co-author, *Factors in Judgment of Writing Ability.*

DONALD D. DURRELL, Professor of Education, Boston University; author, various studies of beginning reading; member, Research Advisory Council, Cooperative Research Branch, U. S. Office of Education.

WILLIAM H. EVANS, Associate Professor of Education, University of Illinois.

EDMUND J. FARRELL, Supervisor of the Teaching of English, University of California, Berkeley.

GARLIE A. FOREHAND, Associate Professor of Psychology, Carnegie Institute of Technology.

N. L. GAGE, Professor of Education and Psychology, Stanford University; past president, American Educational Research Association; general editor, *Handbook of Research on Teaching.*

PATRICK J. GROFF, Associate Professor of Education, San Diego State College.

CHESTER W. HARRIS, Professor of Educational Psychology, University of Wisconsin; past president, American Educational Research Association.

THEODORE HARRIS, Professor of Education, University of Wisconsin; author, books in educational psychology and studies in teaching language arts.

FRANCIS A. J. IANNI, Director, Cooperative Research Branch, United States Office of Education.

LEWIS LEARY, Professor of English and Chairman, Department of English, Columbia University; editor, *Contemporary Literary Scholarship;* author, various critical studies.

ALBERT H. MARCKWARDT, Professor of English and Linguistics, Princeton University; member, Commission on the English Language, National Council of Teachers of English; author, *American English;* co-author, *Facts about Current English Usage.*

HENRY C. MECKEL, Professor of Education and English, San Jose State College; author, chapter on research in composition and literature, *Handbook of Research on Teaching.*

HAROLD E. MITZEL, Assistant Dean, College of Education, Pennsylvania State University; co-author, chapter on classroom observation, *Handbook of Research on Teaching.*

EDWIN A. PEEL, Professor of Psychology, University of Birmingham; 1963-1964 fellow at the Center for Advanced Study in the Behavioral Sciences.

LOUISE M. ROSENBLATT, Professor of English, School of Education, New York University; director, 1963 research development seminar on English; author, *Literature as Exploration.*

DAVID H. RUSSELL, Professor of Education, University of California, Berkeley; 1962-1963 president, National Council of Teachers of English; past president, American Educational Research Association; author, *Children's Thinking* and studies in reading and language.

WILLIAM D. SHELDON, Professor of Education and Director of the Reading Center, Syracuse University; past president, International Reading Association.

JAMES R. SQUIRE, Professor of English, University of Illinois; Executive Secretary, National Council of Teachers of English.

ERWIN R. STEINBERG, Associate Professor of English and Dean of Margaret Morrison College, Carnegie Institute of Technology; 1963-1964 coordinator, Project English, U.S. Office of Education; author, *Needed Research in the Teaching of English.*

ROBERT WRIGHT, Professor of English, Division of American Thought and Language, Michigan State University.

RESEARCH DESIGN AND THE TEACHING OF ENGLISH

I. INTRODUCTORY STATEMENTS

The Purpose of the Conference — David H. Russell 1

Problems of Testing and Research Design in Curriculum Study in English — Erwin R. Steinberg, Garlie A. Forehand 4

Suggestions Emerging from the New York University Research Development Seminar — Louise M. Rosenblatt 12

This Side of Paradigms: The State of Research on Teaching English — N. L. Gage 22

II. RESEARCH METHODS IN VARIOUS AREAS

Research Methods of Literary Scholarship Which May Prove Useful in Educational Research — Lewis Leary 33

Research Methods of the Linguist Which May Be Useful in Educational Research — Albert H. Marckwardt 42

Problems and Possibilities of Research in the Teaching of Written Composition — Paul B. Diederich 52

Research Design in Elementary School Reading and Language — Donald D. Durrell 74

Research Designs Needed in Studying the Teaching of Language and Literature in the Secondary School — Henry C. Meckel 82

III. SOME TECHNICAL APPROACHES

The Place of Systematic Observation in Classroom Research — Harold E. Mitzel 93

Some Recent Developments in Factor Analysis — Chester W. Harris 101

The Analysis of Preferences — E. A. Peel 113

IV. SUMMARIES AND CONCLUDING STATEMENTS

Reports of Study Groups

Research in Teaching Language — Walter T. Petty, Garlie A. Forehand......124

Research in Teaching Literature — Patrick J. Groff, William H. Evans........127

Research in Teaching Composition — Robert L. Wright, Richard Braddock..........129

Research in Teaching Reading — William D. Sheldon, Theodore Clymer..........131

Research in Communication and Verbal Learning — Theodore L. Harris, John B. Carroll..............133

Project English—A Look Ahead — Francis A. J. Ianni........138

Evaluation and Implications — Margaret J. Early, David H. Russell, James R. Squire............145

I. INTRODUCTORY STATEMENTS

The four papers in this section introduced the San Francisco Conference on Research Design and the Teaching of English and now serve to indicate some of the material and problems presented in succeeding sections. Russell's statement expressed hope that the conference would go beyond the results of three previous conferences and suggested specific purposes for the meeting here reported. The Steinberg and Rosenblatt papers in this section further help to set the stage by summarizing the outcomes of two previous conferences. Steinberg and Forehand discuss some of their attempts to select or develop evaluative measures in language, literature, and composition. The Rosenblatt paper underlines the value of bringing together, in seminar or conference, specialists in different disciplines. It emphasizes the worth of small-scale research studies and indicates some of the many possible ways of planning research in the teaching of English. Gage's paper gives an overview of problems of research on the teaching of English with relation to models or paradigms of research.

THE PURPOSE OF THE CONFERENCE

David H. Russell

Ladies and Gentlemen, it is a pleasure to welcome you to the Conference on Research Design and the Teaching of English. As you know, this meeting was planned by some members of the National Council of Teachers of English, supported by the Cooperative Research Branch of the U.S. Office of Education, and sponsored by the University of California. On behalf of these three organizations I can say, to what seems to be a large number of friends and professional acquaintances, that I hope your deliberations will be both pleasant and profitable.

On this day we meet in the shadow of the tragedy of President Kennedy's untimely death. We are part of a bereaved nation and world. This is not the time, however, to express our grief or our admiration. At the opening meeting of the convention of the National Council of Teachers of English, a tribute will be paid to the memory of a man who not only consistently exhibited concern for human welfare, but who contributed greatly to scholarship and education.

Patterns for Research

Without imposing my own ideas, may I now briefly suggest two or three purposes of this conference. As some of you know, it is the fourth in a sequence of meetings which began with one held at the Carnegie Institute of Technology a year and a half ago. That first meeting, directed by Erwin R. Steinberg, was followed by a seminar at New York University, chaired by Louise M. Rosenblatt, and by a conference of chairmen of English Departments at Allerton House of the University of Illinois, directed by Robert W. Rogers. Because of these previous meetings there is one thing we know we do *not* want from this conference: We are not interested in another *list* of needed studies in the teaching of English. I ask you to engage in an even more difficult enterprise, that of exploring in some depth the process of investigating problems in English. We want some patterns, if patterns are possible, for empirical and other approaches to the study of English teaching. One of our purposes, accordingly, is that of producing some models, some suggestions for typical ways in which problems can be analyzed and attacked. We want some alternatives to the two-group, test-teach-test pattern. We should like to know the possibilities of combined qualitative-quantitative methods such as content analysis. It may even be possible for a group here to propose one or two new research tools or devices.

Another purpose is to relate various methods of research to the diverse tasks of the English teacher—tasks which range from having students write a decent paragraph to having them develop critical abilities in literature. These different tasks demand varied abilities, diverse methods of teaching, and, therefore, different methods of experimentation. Although we seek no single road to experimental success, we also must avoid wide scattering. One of the important problems of this conference, accordingly, is that of reconciling and perhaps combining the research methods of the literary scholar and those of the applied scientist. Interdisciplinary is still a fashionable word. Does it have meaning for research in the teaching of English? I was bold enough to say at a conference last spring, "It is high time we had more facts about the teaching of English whether derived from literary scholarship, psychological research, or, even better, a combination of the two approaches."

I hope this conference, including its section meetings, will not be jockeyed into some sort of either-or, black or white position regarding quantitative and qualitative approaches to our problems. Rather, it is my hope that one outcome of your deliberations will be suggestions for cooperative attack on problems, by scholars and researchers in several fields, combining in different amounts the different ways of seeking solutions to problems.

Theories of English Instruction

I make my next suggestion with some hesitation. I wonder if it is asking too much of this conference to give us some leads toward a theory, or theories, of English learning and instruction. I hesitate because theory-making, due to the influence of the natural sciences, is perhaps given too much emphasis in many disciplines today. (If a young professor at any university wants rapid promotion, he should develop a theory.) Despite dangers, I believe we should move in this direction of English. We should clarify basic terminology; we should move away from data accumulation and the merely descriptive to develop a theoretical framework within which varied and large-scale research must operate. Could the conference, or part of it, stress, then, a few examples of deductive thinking, of models which might lead to the building of theoretical systems? I am not sure that attaching these to a hypothetical concept of *structure* in English will be as helpful as relating them to learning theories, but I leave that to some of you who know more about it than I.

Finally, I hope the conference will make a few suggestions about ways of working. If we may be so bold as to assume that gnawing problems of instruction bother English teachers, and that faculty and graduate students in both English and education departments have worthwhile problems in mind, at what stage does the expert in research design and/or statistics enter to help? Should he suggest problems in the very beginning that seem to his unbiased eye to be the important ones? Should he be in on preliminary discussions that attempt to clarify the problem, or should he come in later to review rather complete proposals made by others? Should he help set up clear-cut procedures for collecting data, or should he be ready to adapt methods of statistical inference when the data are partly collected and their characteristics better known? What is the role of the design person in relation to prospective and actual research workers, whether in English, linguistics, psychology, education, sociology, or other disciplines related to the teaching of English?

In the eight or ten hours you have for study together, you will want to react to the papers presented here. But in addition, I hope you can tackle the questions of model studies, varied methods of research, interdisciplinary approaches, and the role of the design expert. And that may be enough for one conference.

PROBLEMS OF TESTING AND RESEARCH DESIGN IN CURRICULUM STUDY IN ENGLISH

Erwin R. Steinberg and Garlie A. Forehand

In the two years since Project English was initiated, many people trained to do research in language and literature have struggled with another kind of research: the evaluation of particular methods of teaching, of new courses, or of experimental curricula. The reason is simple enough. When funds are expended, there should be some evaluation of the worth of the project.

New teaching methods, courses, or curricula should have something more to recommend them than the enthusiasm of their begetters, or the testimony of converts to new programs. The best safeguard is to ask for some evidence of how well a new program works. And such a demand requires evaluation of that program.

So far, of course, we have said little that any of you did not know. Even if it means stating the obvious, however, it is important here to declare the principle: If we want people to adopt the new approaches to teaching English developed under Project English, we must make it possible for them to see the difference in effect between those approaches and the ones we are asking them to renounce. And the burden of proof here is on the innovators. *Caveat emptor* may, perhaps, be appropriate for the marketplace, but it certainly will not do for the school.

Unfortunately, few of us who teach language and literature at any level are equipped either to lay out or to carry out the kind of experimental design necessary to evaluate a course or curriculum. This lack of training frequently has one of two regrettable effects: Either the English teacher is discouraged and decides not to prepare a proposal, or he prepares one, and its research design is so weak that the proposal is rejected.

Fortunately, many English teachers have recognized another alternative—consulting a psychometrician. Many proposals coming in to Project English are being submitted jointly by English teachers and psychologists. The reason that more such joint proposals are not being submitted is probably that it does not occur to the English teacher with an idea for a project that a psychologist would be interested in it. But the psychologists are indeed interested, as the involvement of many of them in Project English studies shows. We are opening new areas of research which present difficult problems of evaluation, and difficult problems always attract people with good minds.

Partnership with Psychology

There are important advantages to the English teacher in his inviting the psychologist in to partnership early. First, the psychologist can help him define his goals in a particular way before the lines of the investigation become inflexible. "If we are going to evaluate the program," the psychologist will probably say, "you will have to tell me what you are trying to accomplish." Because he must have his answer in behavioral terms ("What do you expect the student to be able to *do* as a result of this lesson that he wouldn't be able to do without it?"), he will push the English teacher to an examination of his procedures and his goals in a way that another English teacher probably would not.

When one English teacher tells another, "When students come out of this course, they know how to read fiction," the second English teacher thinks that he knows what the first one means. The psychologist, however, cannot settle so easily. It is not that he doesn't trust the English teacher, but that he must evaluate in terms of *performance*. What does "knowing how to read fiction" mean? If a student "knows how to read fiction," what will he do when he reads fiction that a student who doesn't know how to read fiction will not do? Understanding the answers to these questions, by the way, makes not only for better evaluation but also for better teaching of English.

Precisely these questions too few English teachers have asked themselves. Many of them have theories about what the teacher should do to teach the student how to read fiction, but few of them know, in behavioral terms, what they want the student to do as he reads. Yet how are they going to train competent readers unless they do know? By asking for behavioral answers to the questions he asks, therefore, the psychologist causes the English teacher to consider a dimension of his teaching of which he may not have been aware before. He asks different kinds of questions.

By making a psychologist a partner in a project rather than bringing him in after it has been launched, the English teacher gains in other ways. He is sure, for example, that the proposal will be designed for maximum usefulness. Too often when a psychologist is asked to evaluate data already gathered, he finds that he cannot do so because they have not been gathered correctly. It may seem elementary to suggest that a psychologist be asked to consider the research design when the proposal is being written, but, unfortunately, many proposals come in with evaluation sections which say, in effect, "After we have gathered the data, we will hire a psychologist to analyze them."

A psychologist can also be responsible for making sure that the resources are at hand for carrying through the required research. He knows, for example, what tests are available and how good they are—or he can find out quickly. He knows what automatic data processing equipment is needed and whether it is available. He can work with public school administrators, counselors, and teachers to organize the testing program in a school or schools and give the tests. In short, he can do well many things that the English teacher may have little interest in or experience with.

Contrast and Outcome

The relevance of the experimental design to the aims of the project as a whole rests upon continuous communication between the course designer and the psychologist. The subject matter specialist can facilitate this communication by understanding how a psychologist thinks about research. A psychologist planning a research project will almost invariably try to accomplish two things. First, he will try to define a *contrast* in the way individuals are treated: two groups treated differently, for example, or one group treated differently at different times. Secondly, as we have said above, he will try to define outcomes in terms of observable or potentially observable behavior. These strategies, of course, are reasonable and are consistent with the objective of discovering what happens to students undergoing a learning experience. But they represent a way of thinking about the project that is different from that of the course designer, whose focus is on course content. The course developers may say that the new course or curriculum should be demonstrably superior to "the traditional method" with respect to the student's "perceptiveness" or "sensitivity." But somehow this vague definition of the contrast and the objectives must be translated into operational language; if the psychologist is left to his own devices in making the translation, the risk of misinterpretation and uninterpretability is great. The task of specifying the contrast and the outcome is one on which the English teacher and the psychologist can and must work creatively together. Neither can do the job alone.

In this kind of partnership, both parties accept responsibilities, and the psychologist has as much to learn as the English teacher. A psychologist may be highly skilled in research design and testing but know little about educational objectives or about the operation of school systems. An experimental design, however sophisticated, will be unenlightening if it mechanically contrasts an experimental class whose theory the experimenter doesn't understand with a control class whose content he knows nothing about. If the behaviors that the psychologist chooses to observe are unrelated to the intended accomplishments of the educator, the effort will be wasted. It is not sufficient for the psychologist to proceed from the verbalized objectives of the project; he must understand those objectives.

Thus, experience in Project English has shown that bringing a psychologist into a research project as a partner helps a would-be Project English group to avoid many problems of research design that would otherwise occur. Bringing in the psychologist is the solution to the problem of research design which arises most clearly from reading the proposals for studies and from listening to the advisory panels evaluate them. It is also the solution to the problem of how teachers of language and literature, traditionally without training in experimental design, undertake the evaluation necessary in Project English.[1]

So far we have addressed ourselves only to the general problem of cooperation. Now let us become a little more operational and discuss some of the specific problems faced in the evaluation of a curriculum study. Since we are more familiar with what is happening at the Curriculum Study Center at Carnegie Institute of Technology than at any other, we will draw our examples from there; but the problems we discuss probably are typical of those encountered at any Curriculum Study Center.

Our Project English Center at Carnegie Tech represents the work of four members of the English department, a group of high school English teachers, and one resolute psychologist. For the past year and a half we have struggled with the problems of designing research that is at once humanistically, educationally, and behaviorally oriented. The experience has been sometimes frustrating, often exciting, and always instructive. A glance at some of our problems and attempted solutions may be helpful for others contemplating such research.

Objectives and Rationales

The problem of defining a contrast is a matter of constant discussion in our lesson planning sessions. What makes this lesson unique? Is the approach a completely novel one, or is it better considered a distillation of the best of current practice? If it is novel, what should students learn with it that they wouldn't learn without it? The answers to such questions are recorded as objectives and rationales for each lesson and are subject to further discussion and revision as the lessons are tried out.

---

[1] Erwin R. Steinberg (ed.), **Needed Research in the Teaching of English.** Cooperative Research Monograph No. 11 (Washington: U. S. Government Printing Office, 1963).

Since most of our co-workers are experienced high school English teachers, these discussions help us to understand current practice and thus to make reasonable inferences about the nature of the control classes with which our Project English classes are compared. And, finally, the inferred contrast between the experimental and control classes leads us to predictions about ultimate contrasts in the behavior of the students. We do not always predict, for example, that our students will outperform control group students on any test that might be given them. Rather, we often end up with the expectation that on *some* tests—those that are especially well matched with our objectives—our students ought to excel; but on others, they may well not score as high as control students; and on still others, they ought to perform at least as well as control students if we are to be able to advocate adoption of our curriculum when it is completed.

The careful specification of the rationales and objectives of a lesson has another advantage: It permits the identification of the point in a course at which a student should have mastered a particular concept or attained a certain level of performance. It enables the evaluator to define contrasts in terms of time rather than of competing curricula and thus to study more explicitly the growth of a student's competence.

Motivation and Evaluation

A recurrent and vexing problem in the definition of contrasting groups is that of assuring the equality of the groups in initial ability and motivation. Word of an experimental project seems to travel rapidly, and pressures build up from students, parents, and administrators to get a student in—because the program is new and exciting—or to keep him out—because it is newfangled and may require too much work. An imbalance in student quality brought about by such pressures can negate careful experimental designing. We have approached this problem by working closely with administrators and with guidance counselors, who have the responsibility for selecting student participants. On our first dry run of one of our courses, students assigned to experimental groups tended to be slightly but significantly superior in available ability measures to those assigned to control groups. To avoid a repetition of this imbalance, the guidance counselors and the Project English psychologist worked together to devise a system for assigning students impartially from an eligible "pool." The plan included specifying criteria for eligibility in terms of intelligence and aptitude test scores, past grades, and teachers' ratings, and the development of a guide for assignment based upon random number tables.

The selection and development of outcome measures has also stimulated discussion. We decided early in the project to select tests assessing skills and concepts in the three major curriculum emphases: language, literature, and composition. We learned quickly that a relevant test cannot simply be plucked haphazardly from the many tests whose titles and manuals suggest that they are in accord with our objectives. We ended sometimes by conducting long and intensive searches for a relevant test, sometimes by developing our own, and sometimes by concluding that our greatest contribution could be made by initiating research into the measurement of a kind of performance relevant to our objective.

For example, we examined more than forty tests of English grammar, mechanics, and usage, and felt that none assessed a student's knowledge and understanding of the structure of his language. Some required answers phrased in highly technical vocabulary. Others asked for mechanical identification of "good" usages or for straightforward application of rules. From among all of the tests that we examined, we found one that we were willing to trust to assess the effectiveness of our teaching plan.[2] This test requires that students understand the function that words serve in sentences and thus shares a theoretical base with our curriculum emphasis.

Our search for an appropriate test in literature did not end so easily. The available tests assessed either comprehension of short passages—often emphasizing factual understanding—or the range of one's repertoire of literary allusions (e.g., In whose work does Quasimodo appear?). Eventually we decided to construct our own Literary Discernment Test. Such a test, we decided, should have several properties: It should assess a student's perception of a whole work, his appreciation of what the author has accomplished and how he has accomplished it, and his integration of the parts of the story among themselves and of a story with his own experiences. The meetings, discussions, writing, rewriting, editing, and re-editing that have gone into the attempt to accomplish these purposes have taught us much about the relationship between subject matter content and psychometric properties in a test; they have also resulted in unusual familiarity on the part of five Carnegie Tech faculty members with a certain short story by Akutagawa.

The current version of the test is based on the reading of a complete short story. There are items in three categories. One category includes items requiring critical analysis of the craft of the writer. For example, one item has the stem: "The author tells us that March 3 was a 'clear and cloudless day.' What was accomplished by that description?" The student chooses one of four possible answers. A second category of items refers to the reader's reaction to the story (e.g., "Which of the following events contribute most to the humor of the story?"). The third category proceeds from our definition of literature: "a re-creation in a verbal art-form of what it is like to be alive." These items ask for identification of central themes in the story and for their relation to "what it is like to be alive." (For example, one item stem is "Which of the following quotations most adequately reflects a central theme of the story?")

---

[2] John B. Carroll and Stanley M. Sapon, The Modern Language Aptitude Test (New York: Psychological Corporation, 1959).

The objective of assessing competence in composition met with similar obstructions. Again, no available test seems to measure such skills effectively, and the difficulties of judging student writing consistently and reliably are well known. Our strategy in this instance is to concentrate upon student writings and to work in two directions. First, we try to obtain judges' grades that are as valid and as consistent with our objectives as possible. In one experiment, four judges (Carnegie Tech English professors, none of whom is on the staff of the Curriculum Study Center) rated compositions on five categories (ideas, form, flavor, mechanics, and wording) after discussing objectives at length and after jointly grading and discussing a number of papers. The inter-judge agreement in grades was higher than that often attained in such studies and seems sufficiently high to warrant use of grades thus obtained as evaluation measures. The second direction we have taken in evaluating compositions has been to define as many objectively observable characteristics as we can in a student's writing. For example, we can categorize the topic selections that a student makes when instructions leave partial freedom of selection. Techniques of structural linguistic analysis offer a rich array of untried tools for describing what a student has done in writing a composition. Learning how to select and use these tools meaningfully will, of course, require a great deal of research, and we will probably not be in a position to take full advantage of them in the current Project English study. We have, however, begun some exploratory studies in that direction.

Summary

It would be foolish to claim that we have solved, or can solve, the problems of research design and testing in Project English. On the other hand, we cannot wait for perfect measuring instruments and experimental conditions before evaluating any more than we can wait for perfect curricula before teaching. We must marshal the best evidence we can that curricula we develop under Project English do—or don't do—what we claim. We hope that this discussion of our problems and strategies has illustrated that testing and research design are not static formulas to be mechanically applied but are part and parcel of curriculum research. If the curriculum studies now being so actively developed can achieve a creative integration on subject matter, teaching methods, and evaluation techniques, the result will be advantageous both to the teaching of English and to the behavioral sciences.

SUGGESTIONS EMERGING FROM THE NEW YORK UNIVERSITY RESEARCH DEVELOPMENT SEMINAR

Louise M. Rosenblatt

The Research Development Seminar in the Teaching of English held at New York University from February 27 to March 2, 1963, was a logical step in the series of conferences and publications which David Russell has outlined for us. The project was most directly an outgrowth of the Carnegie Institute of Technology Conference. The proceedings of this conference had provided a broad overview of unanswered questions and recurrent problems in the field of English teaching.[1] After this much needed general conspectus, a valuable next step, it was felt, would permit a small group of scholars and teachers in English to sit down with specialists in research design, psycholinguistics, educational psychology, and tests and measurements, to work together on problems of research in the teaching of English.

Out of the profusion of questions and problems clamoring for solution which the Carnegie Institute report presented, a few specific problems were to be selected, and appropriate research designs were to be devised, in an effort to arrive at some useful ideas concerning research methodology and research design specifically in the field of English teaching. . . .

The planning of the seminar was also influenced by the implications of the resolutions passed at the Allerton Park Conference.[2] The important affirmation that research in teaching and in the instruction of teachers can contribute to the same ends as research in the aspects of English usually considered as part of a liberal discipline threw into relief the need to orient specialists in English toward educational research methods. This view was reflected in the assumption that the results of the seminar should be directed toward some members of the English profession who might not have had prior experience in educational research.

The seminar met as a whole for the first day and a half for exploratory discussions, then divided into groups of at least two or three for work n specific problems. Unlike the earlier research conferences, the seminar did not follow the pattern of presentation of prepared papers. The nembers of the group, in preparation for the meeting, had received opies of the Carnegie proceedings, a statement of the basic aims of the eminar, and a list of guiding questions.

---

[1] Erwin R. Steinberg (ed.), **Needed Research in the Teaching of English.** Cooperative Research Monograph No. 11 (Washington: U.S. Office of Education, 1963).

[2] Richard Wasson (ed.), **Proceedings of the Allerton Park Conference on Research in the Teaching of English** (Urbana, Illinois: Department of English, University of Illinois, 1963).

Improvement of Instruction

One facet of the discussions had to do with the place of research in the current movement for the improvement of instruction in our schools and colleges. Some of the negative attitudes toward research in teaching were faced. There is, for example, the fear that emphasis on research would siphon off energies and funds from such urgently needed practical efforts as curriculum development or teachers' institutes for strengthening the competence of present and prospective teachers of English. Various types of curricular improvement or the development of higher quality textbooks, it has been argued, need not wait on results of future research.

Another problem faced was the general diffidence, perhaps skepticism is a more accurate word, of many members of the English profession toward educational research. This seems to take the form of hostility to quantitative methods when applied to the teaching of English. A frequently expressed fear is that the demand for research problems amenable to quantitative handling would lead to testing for variables that might be easily measurable, but might not be very important from the point of view of the teacher in the classroom. Would not research problems and research designs be tailored to fit the limitations of statistical method rather than the live concerns of the profession itself? Some skeptics, it was said, query whether research projects are not in actuality elaborate means of bolstering what we already know about the teaching of composition or literature.

Reorientation of Profession

In facing the fact that such attitudes are characteristic of many members of the English profession, the seminar members presented the elements of a rationale for research in this field. If a current of productive research in English is to be developed, some process of reorientation, or perhaps reeducation, of the profession seems to be required. The teacher of English has to decide how much his present practice is simply tradition and unchallenged routine; he must recognize that he is usually basing his opinions on personal impressions and criteria. Those who are concerned with creating higher standards or developing new sequences of study or new techniques for teaching must enlarge as much as possible the area of systematic and objective bases for choice of programs and practices. . . .

Out of the reading of the transcripts emerged for me an encouraging view of the place of research in the present drive toward improved instruction in the teaching of English in this country. The whole notion of competition among such efforts as curriculum development or the retraining of teachers and systematic research efforts seemed to fade as the discussion turned toward illuminating the rather complex relationship of these various types of activity. Instead of competing, they reinforce one another and are actually essential to one another as aspects of a massive effort at improving the teaching of English.

Direct Findings on Specific Problems

The major contribution of research should be, of course, its direct findings concerning specific problems. But the effect of fostering research activities was seen as strengthening the other modes of attack. Experimentation would create an incentive toward interest in the possibility of alternative teaching approaches, practices, and materials. Any large-scale research project would involve many teachers and many schools or colleges, and would disseminate information about good practices throughout the country. Moreover, the effort to work out experimental designs and to arrive at some method of measuring results requires a much more stringent thinking through of aims and outcomes and of ways of testing them than is usual in the normal teaching situation. Thus, an experiment can be viewed as also a kind of demonstration and a pedagogical device for disseminating critical thinking about teaching.

Similarly, interest in research requiries support of efforts to improve the competence of people now teaching English. If the profession as a whole provided a richer environment of lively and informed thinking about problems of teaching, probably the calibre and usefulness of research would improve. In an atmosphere in which teachers would be developing new ideas and evaluating traditional and new practices, meaningful research would be more likely to be generated, and the findings of sound research would be more likely to receive serious attention and application.

The emphasis of the seminar members on the need for both basic and applied research should, it seems to me, also be part of any sound rationale of educational research. The assumption was that the profession should move energetically from both ends of the research spectrum—basic and applied. In the most usual sense, "basic" implies "pure" research, free of any concern for future applicability. The seminar members pointed out that some basic research, aimed at knowledge or systematic theory valuable in and for itself, would at the same time yield results valuable as a foundation for more precisely formulated and significant research problems in English teaching.

Hence, an important contribution to the improvement of research would be the strengthening of the theoretical framework on which research depends. Efforts at definition and elucidation of the various outcomes aimed at in "composition" or the teaching of "literature" lead, of course, into problems in rhetoric and the theory of literature. The paucity of strictly defined, theoretically sound research problems reveals also the lack of the kinds of knowledge which generate fruitful hypotheses. Such basic research is needed in all areas–language, composition, reading, and literature.

Central Concepts

Another important area of basic research involves the determination of whether there are certain principles or concepts which are central to the whole field (or to particular segments of the field) of English. A parallel concern is the relationship of language behavior to the mental, emotional, and social development of the individual. From results of these two types of basic inquiry conld flow research directed at the construction of a sequential curriculum, presenting central concepts in a spiral of increasing sophistication and complexity. This would also require research, for example, to determine the levels of linguistic and aesthetic sophistication required of the reader by specific literary works. Moreover, research might answer practical questions about the processes going on in the English classroom: How do young people in our culture come to participate in its literary institutions? What are the operations involved in the inductive and deductive approaches to grammar or to literary criticism? Basic analysis of the interactions through which students are being led both inside and outside the classroom (in relation, for example, to the mass media) could contribute to the sociology of literature. Such analysis would also facilitate setting up more refined controlled situations or correlational studies for testing the effectiveness of teaching techniques.

The group was fertile in suggestions for specific studies that would fall under the rubric of basic research. Time permits my mentioning only a few illustrations. Studies in linguistics represent an area, basic in that it needs no justification outside itself, yet from which may flow significant hypotheses about teaching composition and literature. Examples are a study of the relationship of the writing system of English to its phonology, a longitudinal study of the relationship between the spoken and written language of children through grade 12, or a study of how the individual acquires the language.

Examples in composition would be a study of the development of the English paragraph from 1890 to the present, operational analysis of the production of competent pieces of writing, studies in stylistics, and efforts toward the development of a sound modern rhetoric.

Research and Reading

Much of the basic research in the areas already mentioned would also contribute to improvement in research on the teaching of reading. Other basic research relevant to reading and literature would explore the nature of the reading process as it is carried on with different kinds of literary materials, would study the development of the capacity to understand figurative language, and would investigate the relation between the reader's personality, past experience, and literary preconceptions and various kinds of literary works. The need for a new, or fully developed, theory of literature was stressed.

Recognizing that, until such types of basic research are carried out, we may reinforce aims that are ill-conceived, we must still take the chance and try to fix on those items about which there can be some general, if always tentative, agreement. The seminar members suggested many different kinds and levels of applied research, concerned with evaluating the efficacy of what goes on in the English classroom. More important to us here is the fact that the seminar elicited a number of possibilities in both basic and applied research which introduce a considerable degree of flexibility into what has sometimes seemed a rather rigidly limited field. . . .

The seminar discussions led one to conclude that projects of widely differing scope can be useful at this time. Generalizability of results need not be the sole criterion of the value of a project. Studies small in scope may nevertheless generate hypotheses worthy of large-scale testing. Fairly limited "pilot studies" may clarify the nature of the possible factors at work in a situation and may thus lead to clearer formulation of hypotheses and better organized further research. The same research design might be carried out in projects ranging from the individual classroom to a large-scale effort. The important thing, of course, is that the significance of the findings not be illegitimately extended beyond the limits of the particular research project. The assumption is that the studies of smaller scope would be carefully designed so that they might be replicated or so that they might be effectively modified.

This emphasis on the potential contribution of small-scale investigations was paralleled by a readiness to explore the possibilities for large-scale projects. Rigorously controlled, statistically sophisticated experimental designs, it was suggested, might be carried out on a nationwide scale, perhaps through a network of research centers in English.

The seminar discussions also opened up the possibility of a wide variety of types of research methods. Many of the questions suggested for investigation, and most of the projects outlined, implied, it is true, some type of controlled research design. Difficulties in isolating factors present in the teaching situation and lack of specificity concerning desired outcomes make the planning of such formal experimental research projects extremely difficult. These are, of course, the problems toward which basic research should be directed. Hence, some members of the seminar urged greater utilization of descriptive and correlational studies. Some members of the group envisaged what might be called "quasi-experimental" designs.

Flexibility in Design

Emphasis on flexibility in selection of research design and technique suggested that there are various paths open to English specialists seeking to make a contribution toward a more systematic understanding of the teaching process. Implied also was the need for more general understanding in the English profession of the value and limitations of particular research designs. Understanding of the criteria of sound research is a prerequisite both for wise planning of research and for ensuring discriminating use of its findings. The seminar discussions suggested the possibility that a substratum of informed investigations, varying greatly in scope and method, could have a cumulative effect and could raise the level of the whole research enterprise in English teaching.

Another noteworthy interest of the seminar involved the field of measurement. This problem of the development and evaluation of measurement instruments is, of course, crucial, and requires the imagination and inventiveness of many members of the profession. Here again, much basic research, such as has been suggested in linguistics, stylistics, or the theory of literature, would be relevant to the clarification of what is to be measured.

A survey and critique of measurements in the field of English teaching was one of the major proposals made during the seminar. Another suggestion was that a "kit" of measuring instruments might provide some basic tools and guidelines for the teacher of English both in the classroom and in the experimental research situation.

These, then, were some of the major lines of thought that emerged during the first day and a half of exploratory discussions. After that, the seminar divided into small groups in order to work on plans for hypothetical research projects in the teaching of English. Within the brief period of two or three sessions, about a day's consultation, each group selected a question, developed a common understanding of its meaning, and worked out a proposed research approach to it.

The value of such interdisciplinary collaboration need not labored here. In this seminar and in the earlier conferences, the importance of bringing in research design specialists at an early phase of the planning was stressed, not in order to tailor the design to their prior requirements, but rather to make sure that the final data will be interpretable.

Definition of Terms

One aspect of this kind of collaboration seemed to me especially salutary. Terms which are very much taken for granted by the expert in his own field had to be defined for the person in the other discipline. It was important to spell out operationally exactly what goes on in the English classroom in the teaching of many skills and knowledges which the research expert knows only in the finished product. This was an important stage in the effort to think through a relevant technique for testing the soundness of the teaching method.

When a group felt that it had progressed as far as it could with its project within the limits of the seminar period, it submitted its document, and the members either entered some other group or constituted a new one. When the seminar met as a whole on the last day, fourteen documents were assembled. In the context of the general seminar discussions, they point toward further interdisciplinary work on problems of the methodology of research in the teaching of English.

These documents reflect some of the emphases that I have reported. Some of the projects would lend themselves equally well to small or large-scale treatment. Plans include studies in grammar, composition, the relationship of spoken and written language, the effect of oral reading of literary works on the development of various skills and attitudes in elementary pupils, and the effectiveness of conference or class. One group attempted to approach the problem of the analysis of basic concepts. Four of the documents deal with the problems of evaluation and development of measurement instruments. A list of the titles of the documents follows:

1. A New Attack on the Effectiveness of the Teaching of Grammar in the Improvement of Writing
2. Grammatical Knowledge and Writing Ability

3. An Exploratory Study of the Effects of Various Types of "Pre-Writing Teaching" Techniques on Writing Success of Junior High and High School Students
4. A Proposed Two-Year Study of the Relative Effectiveness of a Rhetorical as Opposed to a Literary Emphasis in Freshman Composition Courses
5. Research Problems in the Relationship of Spoken and Written Language
6. A Study of the Effect of Oral Reading of Literary Works on the Development of Various Skills and Attitudes in Elementary Pupils When They Are First Beginning to Read
7. An Experiment to Determine the Effect of Oral Presentation upon Improvement in Written Composition
8. A Study to Determine the Effects of Wide vs. Limited Independent Reading upon a Variety of Cognitive Processes and Abilities
9. How Can We Measure the Effectiveness of Conference or Class?
10. An Approach to Analysis of Basic Concepts
11. Survey and Critique of Measurements in the Field of English Teaching
12. Model 1 for Measurement of Growth in Writing
13. Toward a Measurement of Growth in Literary Comprehension
14. Inventories of Mechanical Errors

My effort has been to cull from the recorded sessions of the Research Development Seminar ideas or suggestions that might bear on the larger problem of developing a current of significant, cumulatively valuable research, that would, moreover, be assimilated into the thinking and practice of the English profession. . . .

Interdisciplinary Collaboration

Perhaps the moment has come for some of the "further personal comments" that our chairman suggested that I should offer. First of all, it seems to me that this is a particularly favorable time for interdisciplinary collaboration, given the current elaboration of varied and sophisticated techniques by the statisticians and research design specialists. The contributions of the members of the New York University Seminar, like the recently published *Handbook of Research on Teaching*, edited by N. L. Gage, demonstrate the flexibility and insight of the leaders in the field of research design in teaching.

In planning the NYU Seminar, we had intended to invite the cooperation of an anthropologist, but, largely for practical reasons, we did not. Still, here is an area of interdisciplinary collaboration that should be explored. Workers in these fields are often not able to utilize methods of controlled experimental research and are thus forced to devise other techniques of investigation. Their task of seeking some way of isolating and studying specific variables within a complex live situation might, it seems to me, have implications for the study of the complex human situation in the English classroom. Despite the fact that anthropology and the other social sciences have by no means fully solved their own methodological problems, their collaboration might provide an additional stimulus toward the invention of new techniques for discovering indices of the kinds of values and outcomes that we desire in the English classroom. New directions are suggested also by what psychologists themselves are doing in the study of creativity and intelligence and, for example, in recent studies in the selective or planning aspects of communication.[3]

My purpose is not to deny the importance and necessity of experimental research but, rather, to urge a number of different kinds of attacks on our problems. An excerpt from a speech by J. Robert Oppenheimer expresses what I am groping for. Praising the work of Jean Piaget in an address to the American Psychological Association, Oppenheimer made a "plea for a plural approach to exploration, [a] plea for a minimal definition of objectivity."

> I would make this very strong plea for pluralism with regard to methods, that, in the necessarily early stages of sorting out an immensely vast experience, may be fruitful and may be helpful. They may be helpful not so much for attaining objectivity, nor for a quest for certitude which will never be quite completely attained. But there is a place for the use of naturalistic methods, the use of descriptive methods.[4]

In order to benefit fully from interdisciplinary collaboration, we in the field of English need to do a great deal of "homework." Not only do we need the clarification of aims, methods, and outcomes discussed under the rubric of basic research. We need also to meet the challenge to imagination and inventiveness in devising new methods of evaluation and new techniques of research in our teaching field. I am not thinking so much of the application of our formal techniques of scholarly research —literary history or the study of literary techniques and genres. I am wondering, rather, whether we cannot derive much from the processes implicit in literary interpretation itself. For example, the sensitivities, on the one hand, and the checks on subjectivity, on the other, involved in studying the poem and the development of the poet surely are not irrelevant. Such sensitivity and capacity for delicate analysis could be brought to bear on the study of our students as they acquire the English language, the spoken and written word, and on the study of ourselves as we induct them into its benefits in literature. While we seek to carry on the kind of conceptual analysis that will make possible the increasing use of controlled experimental design, should we as a profession not be also drawing on our own particular training for the development of new techniques and new approaches to research on the qualitative aspects of our teaching?

---

[3] Cf. Jacob W. Getzels and Philip W. Jackson, **Creativity and Intelligence** (New York: John Wiley & Sons, 1962); and George A. Miller, Eugene Galanter, and Karl H. Pribram, **Plans and the Structure of Behavior** (New York: Holt, Rinehart & Winston, 1960).

[4] J. Robert Oppenheimer, **Analogy in Science**, presented at the Annual Meeting of the American Psychological Association, San Francisco, California, September 4, 1955. (Mimeographed), p. 12.

THIS SIDE OF PARADIGMS: THE STATE OF RESEARCH ON TEACHING ENGLISH

N. L. Gage[1]

This paper will deal with ideas that are not well understood, that may lead to nothing useful, that may be altogether too ineffable. These ideas seem worthwhile, nonetheless, because they may prove in the long run to be at least as important as the relatively concrete ideas connoted by the term "designs for research." What I have in mind is a discussion of paradigms.

Although the concept of paradigm has been around for a long time in discussions of science, it has recently received fresh attention from historians and philosophers of science. In particular, Thomas Kuhn, in a 1962 volume entitled *The Structure of Scientific Revolutions,* has made much of the importance of paradigms. By one of his definitions, a paradigm consists of the "universally recognized scientific achievements that for a time provide model problems and solutions to a community of practitioners" of science.[2] Paradigms include "ways of seeing the world," "conceptual boxes," "a set of commitments" which determine the problems available for scientific scrutiny and provide "an intertwined body of theoretical and methodological belief."[3] In science, successful new paradigms make old ones disappear and compel the allegiance of all acceptable scientists. New paradigms lead to new journals, new societies, and new curricula for the training of scientists. Having a good paradigm makes it unnecessary for scientists to examine first principles and to justify themselves continually. It is symptomatic of the state of research on the teaching of English, as I see it, that we should reexamine our first principles.

What makes a paradigm successful, and how can we tell when we need a new one? A good paradigm solves crucial problems, eliminates anomalies, and improves the match between the facts and the paradigm's predictions. Normal science, which is what Kuhn calls the activity of scientists working with successful paradigms, consists mostly of "mopping up operations,"[4] and it is what engages most scientists throughout their careers. This mopping up consists of determining facts with greater precision in a larger variety of situations, verifying new predictions from the paradigm theory, and resolving ambiguities through articulation of theory. The paradigm provides criteria for choosing scientific problems. Embracing conceptual, theoretical, instrumental, and methodological commitments "in a strong network," a paradigm sets the task for a whole generation or more of scientific workers in the field to which it applies. Eventually, however, new findings, anomalies, and disappointments of paradigm-induced expectations create a crisis in the science, and a new paradigm appears. Such an event, according to Kuhn, constitutes a scientific revolution.[5]

---

[1] The author is grateful to William Iverson for helpful editorial suggestions.

[2] Thomas S. Kuhn, The Structure of Scientific Revolutions (Chicago: University of Chicago Press, 1962), p. x.

[3] Ibid., pp. 4, 5, 6, and 17.

For his examples, Kuhn draws entirely upon the history of the natural sciences. He gives much detail on the course of the scientific revolutions that yielded such new paradigms as Copernican theory, Lavoisier's theory of oxidation, Newton's laws, Darwin's theory of evolution, Einstein's theory of relativity, and Marie Curie's theory of radioactivity. He might also have mentioned the Mendelian theory of heredity, and in contemporary science such theories as those of antimatter, the Doppler shift, and DNA.

But I am not competent to illustrate Kuhn's conception of the paradigm in these fields. Let me turn instead to ideas with which I am more at home. In psychology the whole field of factor analysis exemplifies a paradigm, originated by Charles Spearman and developed by such men as L. L. Thurstone, Truman Kelley, Godfrey Thomson, and Harold Hotelling. Conditioning theory, as originated by Pavlov and Thorndike, constitutes a paradigm. And, of course, so does the kind of theory associated with the name of Freud.

What do these diverse movements in psychology have in common with one another and with the paradigms of the physical and biological sciences? Each of them has set the pattern for a host of conceptual, theoretical, and methodological developments that has kept several generations of scientists busy in the kind of mopping up operation that, in Kuhn's view, constitutes the work of normal science. Factor analysis is still very much alive today as a way of providing structure and dimensions not just for the realm of mental abilities, with which it got started, but also for attitudes, individuals, and occasions. It has generated new journals, courses, and societies of research workers who share this paradigm and shape their careers according to it. The same things can be said of conditioning theory, still vigorously alive in the work of Hullians and Skinnerians, with their own journals, courses, and societies. And the Freudian theory has led to projective testing, with its TAT and Rorschach and also its own journals, courses, and societies.

---

[4] Ibid., p. 23.

[5] A circularity in the argument of this conception has been noted by Norwood B. Hanson, Professor of Philosophy at Yale. He stated, in a discussion at the Phi Delta Kappa Symposium on Education and the Structure of Knowledge, University of Illinois, November 4, 1963, that Kuhn seems to define paradigms as sets of ideas whose overthrow occasions scientific revolutions and to define the latter as events in science that follow upon new paradigms.

It is somewhat more difficult for me to think of paradigms in the field of scientific work in education. But perhaps the paradigm of evaluation, developed by Ralph Tyler, qualifies as such a system of ideas. As will be recalled, Tyler emphasized the importance of defining educational objectives in behavioral terms as the first step in evaluating educational outcomes. It seems fair to say that a whole generation of measurement workers in education has devoted itself to the application, refinement, extension, and elaboration of this paradigm enunciated and promulgated by Tyler in the early 1930's.

Three Kinds of Paradigms

Now, what has all this got to do with research on teaching English? Do we have paradigms in this field? Does it do any good to talk about them, to get self-conscious about them? Or is it better to leave them, if they exist, submerged in the unconscious of the field—unstated, unexamined, and unchallenged? My answer is that we should get them out in the open, expose them to rational scrutiny, and see if we can do anything to improve them. It may well be that we have nothing in this field at present by name of paradigm. That is, we may have nothing comparable to the powerful systems of ideas that Kuhn dealt with in the natural sciences. We may well be in a pre-paradigmatic stage of development where no scientific achievements have yet compelled the kind of assent that marshals a whole army of research workers behind it. But if we do have paradigms, their state of health badly needs examination. Making this examination may hasten the arrival of better paradigms.

To organize my discussion, I want to divide the overall notion of paradigm into three parts: substantive, methodological, and logistical. By "substantive," I mean the concepts, the educational and psychological variables, and their relationships, that constitute the subject matter of the field, its way of abstracting and focusing on a certain set of the infinite variety of phenomena in the natural world. For example, methods of teaching and types of achievement are substantive concerns of research on teaching.

By "methodological," I mean the ways of measuring variables, and determining relationships between them, namely, the tests and statistical methods, the ways of collecting and analyzing data, that are used in research. In research on teaching, methodological concerns deal with schemes of observation, rating, judging, testing, experimenting, and with ways of correlating, analyzing variance, factor analyzing, and so on.

Finally, by "logistical," I mean the whole host of concerns with men and money, organization and administration, that are entailed in getting research planned, executed, interpreted, communicated, applied, and improved.

These three kinds of paradigms, or parts of an overall paradigm for research, do not, of course, exist in isolation. They interact; each affects the others. Substantive paradigms interact with methodological ones, and these in turn have implications for the kinds of logistics that will facilitate research. Kuhn was right in viewing paradigms as tightly intertwined systems of concepts, theories, instruments, and methods, and my separate treatment of them does violence to the realities of research life. I shall discuss them in this way only for the sake of reducing complication.

Substantive Paradigms

Let me turn now to the substantive paradigms that prevail in research on teaching and, insofar as I have the required knowledge, in research on teaching English. What are the variables and the theories of learning and teaching that have shaped research on teaching? The answer must be put in general terms because paradigms refer to the large conceptions that govern scientific effort and not to the specific forms they take in any given research project. Further, I do not have the background for more than an illustrative attempt to identify these substantive paradigms.

As I see it, current paradigms for research on teaching English have as one of their bases a distinction between composition and literature, between the production of writing and the assimilation, or appreciation, of writing. Within each of these major rubrics of concern, further distinctions seem to prevail. Within composition, influential distinctions are made between various kinds of writing, such as the expository or the literary. Further distinctions are made between levels of analysis of writing—i.e., between the sentence, paragraph, and the whole theme, between grammar and usage, between clarity and emotive power, between denotation and connotation.

I am not qualified to push this kind of identification of influential distinctions much further, even assuming I have not already made some basic errors. Let me carry this effort just a little distance, however, into the field of literature. Here I see research being guided by distinctions between intellectual comprehension and emotional appreciation, between plot and characterization, between prose and poetry, between novel and play, and so on.

These kinds of distinctions provide the bases for the kinds of definitions of achievement that teachers of English concern themselves with. And these definitions in turn underlie the kinds of dependent variables that get involved in research on teaching English. These distinctions, or more accurate characterizations of them than I am capable of giving, constitute part of the substantive paradigms that I think pervade research on teaching English. And the obvious question is, do these distinctions generate fruitful paradigms?

Now let me go to the other side of the classroom, away from what we look for in pupils, toward the teacher, his behaviors and characteristics. What kinds of distinctions and variables, underlying what kinds of paradigms, do we find here? Again, I can offer here merely some examples of what I have in mind. Much research on teaching has been concerned with teachers' traits, such as intelligence, knowledge of subject, dominance, emotional adjustment, and authoritarianism. Much research has been concerned with how much of various kinds of talking and other social behavior the teacher emits and elicits from his students. A smaller amount of research has dealt with intellectual, logical, and cognitive dimensions of the discourse that teachers engage in and draw out of their students. And a still smaller amount of research has dealt with the question whether the effects of teacher behavior vary as a function of the specific personality traits of the pupils with whom the teacher interacts.

The kinds of variables that are studied arise from the implicit theories of learning and teaching held by the research worker as part of the paradigm that governs his effort. If he is oriented toward a conditioning theory of learning, he is apt to study the reinforcements, the rewards and punishments, that teachers distribute to pupils as a consequence of their behavior. His observation scheme, his rating scale, his experimental teaching method are conceived along lines that focus on the pleasantness or unpleasantness of what teachers do to pupils who have behaved in various ways.

It should be noted that the Skinnerian part of the programed instruction movement follows an operant conditioning paradigm of learning and teaching. Hence the corresponding research focuses on problems of reinforcement, stimulus discrimination, and shaping. The latter concept refers to building new responses on the basis of previous ones by reinforcing successively closer approximations to the desired response. Whether such a paradigm can ever be adequate to shape up response chains as complex as good English composition is as yet unknown. Resnick's analysis of the issues led her to an optimistic outlook.[6]

---

[6] Lauren Resnick, "Programmed Instruction and the Teaching of Complex Intellectual Skills: Problems and Prospects," Harvard Education Review, 33 (1963), 439-471.

If the research worker is oriented toward an identification-imitation theory of learning, his research variables deal with the pupil's attitude toward the teacher as a model whose behavior, observed by pupils, will be learned if the pupils like or respect the teacher enough.[7]

If the research worker is imbued, wittingly or unwittingly, with a cognitive theory of learning and teaching, his variables, and the substantive paradigm of his research, will be saturated with regard for the logical, structural, and organizational aspects of the content of what is communicated to and from pupils.[8]

Which of these three orientations toward learning and teaching—conditioning, identification, and cognitive theory—will be adopted by research workers depends to some extent on the subject matter. Motor skills, emotional attitudes, and intellectual comprehension differ in their susceptibility to these three kinds of learning theory, but many workers try to apply a single kind of theory to all kinds of things to be learned.

Now, these paradigms of learning and teaching, or others like them, seem to me to shape the research that gets done, even when the investigator is unaware of them. If the research worker has a successful paradigm, it yields positive results, and we witness a steady accretion in our power to understand, predict, and improve the kind of learning that teachers bring about. But if it is not a successful paradigm, then it ought to be examined.

Methodological Paradigms

In turning now to methodological paradigms, I shall not deal with the technical aspects of matters like experimental design, factor analysis, and scaling. Rather, I want to call attention to the methodological paradigms underlying experiments on teaching methods or correlational studies of teachers' personality traits.

Consider one aspect of the paradigm underlying experiments on teaching methods. Such experiments proceed on the faith that what a teacher does makes a difference that has considerable generality. The generality is usually assumed implicitly to extend over a broad range of subject matters, or at least over a broad range of topics or units within subject matters. Thus it is considered to be general over the teaching of English composition as a whole, including punctuation, capitalization, paragraphing, organization, and appropriateness of style. The teaching method is also considered to have effects that are general over a wide variety of students, bright and dull, middle class and lower class, secure and insecure. The generality of the method is also considered to extend over all the concomitant experiences of students in other classes where they study other subjects under other teachers.

---

[7] A. Bandura, "Social Learning through Imitation," in M. R. Jones (ed.), Nebraska Symposium on Motivation (Lincoln: University of Nebraska Press, 1962), pp. 211-268.

[8] D. P. Ausubel, Psychology of Meaningful Verbal Learning (New York: Grune and Stratton, 1963).

Now, since methods experiments seldom have yielded big differences, it may be that this assumption of generality is unwarranted. The effects of a teaching method may be too much a function of the facet of the subject matter, the kinds of pupils, and the concomitant experiences. Hence these effects may be too hard to detect under the methodological paradigm of generality of effect. A change in the paradigm would call for much more complex multivariate experimental designs that would reveal the significant interactions between teaching methods, subject matter components, pupil characteristics, and concomitant experiences. Teaching methods, such as critical versus tolerant grading of themes, would not be studied in isolation. Subject matter components (such as the use of strong verbs), pupil characteristics (such as ego strength), and concomitant experiences (success in other courses) would be studied in interaction with teaching materials.

A second example of how a methodological paradigm can shape research may be found in the typical correlational study. In such a study, measures of a personality trait in teachers are correlated with measures of a criterion, such as how much pupils have changed in knowledge or attitude. This kind of study follows a paradigm that has had moderate success in industrial and military personnel research. This paradigm has employed two kinds of variables— a criterion of success on the job and a measure of a potential correlate of that criterion. After obtaining measures of the criterion and the potential correlate for a sample of workers, the investigator computes the correlation between the two.

Whatever success this paradigm may have had in military and industrial settings, it has not yielded much in research on teachers. Most findings have been negative and inconsistent. Perhaps the reason may be found in the methodological paradigm which dictates that a major source of variance in success is the person tested. When the person works with inanimate material, as do radio tube assemblers or airplane pilots, his success may indeed be a function largely of his own characteristics. But when the person works with other persons, as do teachers, his success may be a function not only of his own characteristics but of his pupils' characteristics as well, and of the interaction or relationship between his own characteristics and those of his pupils. Pupils are not uniform and inanimate material like radio tubes; they differ from one another, and they bounce back, react, and interact. The typical correlational study may therefore reflect a combination of methodological and substantive paradigms that simply do not accord well enough with reality to yield positive findings in this kind of research. Separate correlations, for subgroups of pupils according to their relevant characteristics, would then be needed. The correlation between a teacher characteristic, such as authoritarianism, and the favorability of pupil response might be positive for insecure pupils and negative for secure ones.

But again my purpose is merely to illustrate the kinds of questions that forthright concern with paradigms can lead to. In the methodological realm, also, paradigms exist, are important, and may become better if we look askance at them hard and long.

Logistical Paradigms

Finally, I wish to apply my interest in paradigms to the logistical problems of research on teaching English. Here, as I have noted, we grapple with the problems of how to get men and money into the social and organizational setups that are most likely to be conducive to the kind of research we need.

Until the nineteenth century, the logistical paradigms of research were simple. Gentlemen—like Newton, Harvey, Cavendish, Priestley, Franklin, and Lavoisier—with the necessary leisure, brains, and curiosity cultivated science. They built the foundations of chemistry, physics, and biology. In the nineteenth century, the modern university came into being, and scientists tended to be professors who had their own laboratories, constituted their own departments, taught their students, and did their research. The German universities brought into being the corps of graduate students, who attached themselves to the professor, while the latter continued to be very much the master of his own research destiny. Johns Hopkins brought this logistical paradigm for research to the United States. In this country, although departments grew in size, this paradigm seems to have prevailed until about 1941.

During the war, the university-affiliated but almost autonomous research laboratory came into being. The Radiation Lab at MIT, the Metallurgical Laboratory at Chicago, and the Manhattan District Project at Los Alamos may be regarded as exemplifying whole new logistical paradigms for research and development.

In the meantime, of course, the old-fashioned autonomous professor paradigm had been accompanied by the development of the mechanisms called the scientific society and the professional journal. Officers, prizes, committees, and annual meetings were invented. The National Academy of Science was started in Lincoln's time, but the federal government did not enter the picture forcefully until World War II.

Again, this is not the place to trace these developments more adequately. I am merely trying to illustrate what I mean by logistical concerns and paradigms as a basis for remarks about the logistics of research on the teaching of English.

In this field, the logistical paradigms may still be sufficiently unsettled to allow for rational planning. Many questions arise, and the questions themselves need to be organized along lines of generality and priority. But it may do some good to put some of these questions now.

First, where are the research workers coming from? Is research on *teaching* English sufficiently well established as a career activity to attract the talent it needs? Or is it quite disreputable for graduate students in English to go into such work as a steady thing? And will such persons be permitted, indeed encouraged, to get the training they need in statistics, measurement, and other alien subjects? If not from the field of English, are the needed researchers to be recruited from education? In that event, the fact should be faced that education does not produce nearly enough research workers to staff its own better established opportunities for research careers.

A second kind of question is whether the organizations of teachers of English are now structured in such a way as to foster the proper conduct and communication of research. Are the necessary committees, commissions, divisions, and organizations in existence? Or does the present structure reflect a conception of research on the teaching of English that is inadequate for the scope and variety of the work to be done?

Just one more question: assuming that collaboration between English teachers and research specialists will be needed for some time to come, we can ask whether the paradigms for this collaboration have been thought about. Under what kinds of arrangements can the research specialist do most justice to his responsibilities and opportunities in research on teaching English when his main allegiances and commitments belong to a school of education or a department of psychology and he has already developed all-consuming research interests of his own that may have nothing special to do with English? Is the occasional consultancy or workshop or conference enough? Will the research specialist be brought in only long enough to muddy the water, raise questions, and create anxieties and then be allowed to go home, while the English teachers wonder what they should do next? Or, can more enduring arrangements be erected so that truly fruitful interdisciplinary work will come about? Medicine and chemistry, agriculture and genetics, and industry and operations analysis have worked out such relationships. English and educational research—the two parties we are interested in—ought also to pursue such possibilities of productive alliance. If so, should we not regard conferences like this one as mere introductions preliminary to serious courtship and marriage?

I fear that I am merely demonstrating that a child can ask questions to confound the philosophers. But these questions do refer to what may be the most important kind of paradigm at this juncture in the history of research on teaching English. Although I have not treated them in any orderly fashion, I have tried to sketch the lines along which better questions could be framed. To my mind, research on teaching English will become a serious enterprise only to the extent that solutions to these kinds of problems will be found.

Summary

Beginning with the notion of paradigm as a pattern of concepts, theories, methods, and instruments that govern research, I gave some examples of paradigms from the natural sciences, psychology, and education. Then I divided the subject into substantive, methodological, and logistical paradigms. Applying this orientation to research on the teaching of English, I offered a number of illustrations of what concern with paradigms of these three kinds might look like. Of them all, it seems to me, logistical paradigms are in greatest need of clarification at the present stage of development of research on the teaching of English.

II. RESEARCH METHODS IN VARIOUS AREAS

Because the field of English is a broad one, the five papers in Section II are related to seven broad categories: Leary's to literary scholarship, Marckwardt's to linguistics, Diederich's to composition, Durrell's to the two categories of elementary school reading and language, and Meckel's to the two areas of language and literature in the secondary school. Thus Section II presents a way of looking at research problems and methods according to curriculum content.

Leary believes "that the function of the literary scholar is to tell us what to teach, and that the function of the scholar engaged in educational research is to tell us how it may be taught," but he adds the important note that "neither can speak with complete authority without the concurrence of the other."

Marckwardt similarly contrasts the study of language as language with the investigation of linguistic acts as manipulations of a language system by an individual or a group. He stresses the primacy of spoken language and points out that the teaching and learning of reading is "by no means synonymous with (teaching and) learning language." He asks that the educational psychologist recognize the assumptions and attitudes of the linguist.

Of particular interest to those who want to test specific methods or materials written in composition is Diederich's discussion of reasons why it is difficult to obtain statistically significant results in investigations analyzing changes during instruction. His paper underlines the importance of considering the use of other research designs, a few of which are suggested in Sections III and IV of this report.

Durrell stresses the many questions about learning to read that still await answers and pleads for "a long and complex research campaign" in this area. He believes that the nature of the problem in reading determines the research design. The reader will be interested in his three strategies for improving research and his plans for one of them, a coordinated attack using varied approaches to the topic of initial instruction.

Meckel's paper is a fitting climax to this section because it explores some theoretical problems in the design of studies of language and literature in the secondary school. He uses learning theories as a basis for establishing what he calls a rationale for research in English. Meckel discusses the importance of sequence in the learning of English and concludes with some suggestions for studies in relatively unexplored aspects of the curriculum in literature.

RESEARCH METHODS OF LITERARY SCHOLARSHIP WHICH MAY PROVE USEFUL IN EDUCATIONAL RESEARCH

Lewis Leary

The subject which has been assigned to me seems at the same time exceedingly simple and exasperatingly difficult. Although to one degree or another I have been engaged, though sometimes vicariously, for many years in what has been called literary research, it had not always occurred to me to think that it was in any fundamental way different from research in most other areas; or that it was not subject in every field to the distortions and misdirections and irresistible excesses to which it is subject in my own. I have known that there is some distinction, even in my own work, between things which can be counted and things which cannot be counted; and I have had some sense of being best about my business when I resisted temptations to make equations. In a general way, we have been taught that science and science-like subjects concern themselves with facts of repetition, objectively observed, and obedient—when man is properly learned—to some measure of control; and we have been led to understand that the humanities, especially the arts, rightly concern themselves with matters of succession, less predictable, often apparently uncontrollable.

Sweeping and confident statements such as that one are probably more convenient than true, but to many of my literary colleagues they seem incontrovertible. We are likely to tell you that the function of literary scholarship is to examine and explain the literature which is studied, not to improve it. What improvement is involved is in the taste, the understanding, the experience of the person who will then or later examine the same literature. Insofar as we understand educational research, we think of its purpose as examination in order to effect improvement, not merely to describe and interpret, but to order and reform. The one concentrates on what man has done, so that in the broadest sense its approach is historical. The object of its study is something which has been produced. The other is contemporary. Its object is to endorse or correct an activity which now exists, in order that it may continue to exist at best efficiency.

We are likely, therefore, to suggest that there is little possibility for anything more than friendship between research in literature, which is an art, and research in teaching, which is a social function—the one a complex entity, which begins in the consciousness of the artist and ends in the consciousness of someone else, with what Harry Levin has called the trajectory of the imagination between; the other, an equally complex but man-controlled process, which I know less about and therefore do not define. On one side or the other, we are accustomed to dismiss, and sometimes with arrogance, suggestions that there is also an art of teaching which has not yet yielded to measurement, and that there is a social function of literature which has been variously and probably in many instances not objectively defined. More often than not we in literature are likely to dismiss each proposition, the first as something which cannot be taught and the second as a subject not worthy of scholarly attention. When we leave our carefully cultivated clearings to wander amid the unfamiliar forests of other people's different endeavors, any of us inevitably begins to run uneasily in circles, shouting words which cloud rather than clarify what we have to say. Good fences, we therefore find it convenient to suggest, do make good neighbors.

Because my subject seems to call for equations between methods of research, I am not sure that I can be helpful with any but the simplest and negative, but I hope radical, suggestions. For one thing, literary scholars have not always been known as methodical people or as pioneers in methodology. They have rather been notorious as borrowers of other people's ways of doing things, and not always, it has seemed, to their ultimate advantage. When Taine borrowed the method of the environmental psychologist, many were stimulated, and the study of literature moved beyond accustomed boundaries, sometimes to become confused among a welter of fascinatingly unfamiliar but inappropriate considerations. When I. A. Richards, to offer one more example, applied analytic methods of the educational psychologist, the student of literature learned something more, but less perhaps of the literary object than of the person who reacted to it. Through all her life literary scholarship has been a conscienceless hussy, brazenly parading in garments borrowed from other people, many of whom she has said she did not like or trust at all. Nurtured by theology and philosophy, adopted by history, tempted by science, solicited by sociology, seduced by that foreign rascal Freud, she remains for all that a pretty independent, pert person, flirting still with the way other people do things and continuing to change her costume at the behest of those who make indecent overtures, but keeping through it all something of maiden integrity.

She is easily disconcerted about words, and if she is to any degree a prude, it is in her insistence on their proper use. For that reason many of us who would defend her feel that the word *research* (or *re-*search, as we say those who misunderstand its meaning usually pronounce it)—that word *research* has become a shibboleth among us, so misused, we are apt to say, by pitchmen extolling the excellence of cigarettes or laxatives or pedagogical panaceas that it becomes an embarrassment. Indeed, the Modern Language Association has recently spent hours discussing whether the word should not be dropped from the description of what it as an association of scholars is concerned with. In my university, we no longer use the word research in considering promotion of younger men. We are more likely to use some such phrase as serious and continuing intellectual commitment. Personally, I prefer and customarily use the word *investigation*, as more honestly descriptive of what most of us are about. *Research* sometimes seems now, almost as much as anything else, a word used to tell foundations or government agencies what we want money for. Yet it is an honorable word, worthy of higher repute; we resent and rebel against its misuses. My dictionary, the good dictionary, defines it as "careful search, a close searching; studious inquiry; usually critical and exhaustive investigation or experimentation, having for its aim the revision of accepted conclusions, in the light of newly discovered facts." It is, some of us found ourselves explaining to a student who stood our hair on end with a computer-built dissertation project, what takes place after the IBM machine has done its work—or what is to be done in telling the machine what it must do.

Author, Text, and Book

That is what all of us try to accomplish, bound by limitations of experience and knowledge, each in his own fashion. In literary scholarship we move through three areas, the boundaries of which are not clearly defined, so that one often overlaps another. Like most enterprises which beguile man's inquiring mind, ours in literary investigation forms a familiar trilogy: the creator, the word, and the word incarnate; in our terms, the author, the text, and the book, or studies biographical, textual, and bibliographical, each divided and subdivided and, not at all irresponsibly, encroaching on the others. In spite of warnings that an author's intentions are unimportant, we spend large portions of our time inquiring about the literary creator, where he came from, what makes him act as he does—his ancestral background, his environment, his reading, his tendency to imitate or react against other creators, his attitude toward his mother, the traumatic experiences which have warped his career or made it magnificent.

Our textual studies bring us to knowledge of words, their origins, their usage; of groups of words which have been used in similar combination before or since; of theme and plot and characters, familiar or brightly new; of aesthetic values and moral tone. It is likely to lead us to manuscripts, to the study of variant readings, and even to the study of handwriting or the search for signs of something in the force with which typewriter keys seem to have been struck. Our study of books, a scientific joy to the descriptive bibliographer, leads also to ancillary questions on the economics of authorship or the standards of taste in one generation or another: who made the book, under what conditions, and why; who paid for it, and how much; who read it, and how often, and for what reasons? Was it made to sell, to reflect current taste, or was it a revisionist book meant to lead or expand? Here we circle back on ourselves as study of book subsumes study of author and what he wrote, for as every trinitarian knows, the three elements of any godhead are one, essentially indivisible, and ultimately indefinable.

We do not pretend to be pretentious when we say that the proper area of investigation for the scholar-critic in literature is the best that has been thought and said, and how it has been said, and why that way was (and perhaps still is) the best of all possible ways of saying it. We are concerned with ambiguities and ironies and indirection, with what Robert Frost called a writer's little jokes. We spend much time on things which are not real in any measurable sense, on words and combinations of words which are so reflective and reflecting that they express more than words, even the same words in different combinations, have said before. We are concerned, as William Faulkner once suggested, with active and acting words that lie flat on the ground with relations to things, rather than with words that go straight up in the air, like smoke, dissipated in meaning. For the literary scholar is characteristically a linguist, interested in the romance and the mystery and the possibilities for meanings beyond any meanings we have known before, in words in right arrangements used respectfully and artfully, each in its proper place.

Because the nature of his quest requires him to attempt explanation of what cannot be explained except by indirection, so that the articulation of his best discoveries invites comparison with the literature which it contemplates, the literary scholar is often, despite protective coloration, a modest person. Those among my acquaintances who are not are the people who count and measure, and they seem so useful to the rest of us in unobtrusive ways that we forgive them their trespasses on manners or taste. These are those who may be thought to have most to contribute to scholars in other areas, for they have borrowed methods and sometimes improved them; and the roads on which they confidently travel in rented vehicles do lead, as perhaps all roads do, toward literature, but end abruptly at the walls which counting cannot scale. Properly, at work, the student of literature is a person of informed humility, cautious about stretching beyond his manifest limitations, and therefore manifestly of no use whatsoever to anyone but himself and the very special and very large requirements of the task to which he is dedicated. By choice he remains within the walls, intent on throwing whatever revealing light he can upon the continuously retouched portrait of the spirit of man which is called literature.

Respect for Words

As we have approached this central responsibility of ours, which is aesthetic and meaningful at the same time, except that we cannot prove that except with other words, all of us have borrowed, as I have said, other people's methods of investigation, and with gratitude, though we have been warned by the best among us that they create an undertow which, unless we are very strong, carries us relentlessly away from the study of literature. All that we can offer in return for what we have borrowed is, I think, our concern with words, our respect for their proper use, and our conviction that only through words rightly ordered do we speak sincerely together.

However obvious and iterative and threadbare the charge may be, it must be made again and again—that we do not always speak decently together. Something of snobbishness in use of words stands between us. We wrap them so tightly around what we have to say that meaning is smothered. My colleagues do it, and yours do too. We are not so much secret as we are careless or impolite. The literary man does not always understand the man in education: What in the world, one of my colleagues asked me, are designs for research? And I was not able to answer him. Yet I wish that I could say that one of the things which people explaining research in education could learn from all men who explain literary research is plainness in speaking; but I cannot do that, for some of my best literary friends speak in words which go straight up in the air also.

Our problem, I think, is double. It is first to simplify and clarify, each in his own area, respecting words for what they mean, for their explicit as well as their metaphorical values, not wrenching them out of place, never creating a new word when old words will do. Because so many of the best utterances of men have been simply phrased, we inevitably respond with profound revulsion against linguistic combinations which are decentralized by multisyllabicity. But "when I use a word," said Humpty Dumpty, "it means what I choose it to mean," and so we other eggheads each have our necessary special words which tell of refinements of that of which we speak, and we think we cannot do without them, so that the second requirement upon us is the learning of the other's way of speaking. How terrifying was the statement made some years ago by Robert Oppenheimer, who said that no layman could be expected to understand what the physicist said. Can we suppose that finally no man can speak to another, so private do we become?

Unity among Men

At the conclusion of a conference held some years ago, theologians and physicists and poets and politicians gathered in an attempt to determine whether there was in truth such a thing as unity of knowledge toward which men might cooperatively strive, or whether, as discussion then seemed to imply, man himself was the only unity because of the biological circumstance that he represents a single species of which there are no subgroups. Most of the participants maintained that instead of unity of knowledge there is possible only unity among men who search it out, a unity of spirit which has many names—tolerance, understanding, humility, persistence, devotion, love, in spite of which each generation seems challenged, as much by its ignorance as by what it has learned, to seek out or attempt to manufacture an artificial unity amid diversity, to seek allegiance and help from other men in the work which each man must do alone. At the end of this conference, Archibald MacLeish summed up the consensus by suggesting that what is desirable is not a universe of discourse but a constellation of discourses, not a basic English of all disciplines which would be common to all but adequate to none, but a communication among discourses based on a common understanding of purposes and related to a common spirit of intelligence and humanity.

I am by no means an intellectual isolationist, but I do suggest that in the largest sense we cannot really help each other. My methods are not yours and used to your ends will be distorted beyond usefulness. I cannot afford the distraction you provide. But we can communicate, and we can make trial at understanding, recognizing what is common to our purpose. Yet our best assistance to each other may be assistance to ourselves, each going his own way, doing what by talent and training he is best equipped to do; and then pausing, as in this conference we pause, to see what others have done and to consider the possibility of discovering useful relationships between these things that they have discovered and those other things that we have discovered.

We can talk together, I think, and we can share that which we have in common, which is the circumstance that we are teachers and interested in knowing how best to communicate what is most certainly known or best thought and said. Our experience as teachers is among us a communicable topic, what we have done in this situation or that to make plain a poem or a literary genre or generation or a way of writing essays. Perhaps our meeting tonight in this conference is a kind of research, intelligently calculated and subject to some manner of final estimation, a valuable but, we must think, a limited kind of controlled investigation—disruptive, some may suppose, of that perlucid eleventh commandment, Thou shalt not committee.

We can say this simple thing: that the literary scholar through his professional activities can tell us what is to be taught. Northrop Frye, in reporting the investigation of the Joint Committee of the Toronto Board of Education and the University of Toronto, discovered that the central question in the discussion between these two groups finally asked, "Does the teaching in the schools, or at least in the secondary schools, reflect contemporary conceptions of the subjects being taught?" That is, in our discipline, does the teaching of English keep up with the findings of contemporary literary scholarship? Or do we not come face to face with students who have been extremely well taught, but taught the wrong things? In dealing with the vexed problem of the promotion of our younger colleagues, we have learned that good teaching is by no means the single requirement for academic recognition. With some sense of responsibility for coming to the aid of the beleaguered teacher, so busy with other things that he has no opportunity to read, our NCTE Committee on Literary Scholarship and the Teaching of English has attempted to do some things toward bringing teachers up to date by providing programs at national meetings, and even by producing a book which was just good enough to indicate how much better it might have been;[1] but we have not done as well as we should, and we welcome suggestions for expansion and improvement. That is the responsibility of the literary scholar—to speak plainly to those who do not have his opportunities or talents for investigation, so that all of us may be sure that what we teach is correct.

---

[1] Lewis Leary (ed.), Contemporary Literary Scholarship: A Critical Review (New York: Appleton-Century-Crofts, 1958).

What and How of Teaching

If we say that the function of the literary scholar is, for our purpose, to instruct us in what to teach, can we not then as simply say that the function of educational research is to instruct us in how to teach? Through pride or ignorance, which may be the same thing, at one time or another in the past each of us has turned up his nose at suggestions of the other. (I myself have shuddered at the implications in the situation in which one good friend of mine who is in educational research found himself; he had put together a method of teaching which could teach anything, and he was employed by the War College to teach soldiers how to do just that. That reminds me of the wife of a graduate student in a teachers college who commiserated with my wife when I was also in graduate school because when I received my degree I could teach only English, while her husband could teach anything.) I again underline the obvious when I say that method alone can be disruptive of every value which humane scholarship cherishes. At the same time, it is unnecessary to remind you again of the tragedy of the learned man who has not learned to speak plainly.

Time and Sequence

When we have said these simple things, that the function of the literary scholar is to tell us what to teach, and that the function of the scholar engaged in educational research is to tell us how it may best be taught, and that neither can speak with complete authority without the concurrence of the other; then we may come to a question which as scholars and teachers concerns us all: When is this or that in literature to be taught? We have before us, or shall have when their report is completed, an example of collaboration between educator Arthur Foshay and literary scholar Roy Harvey Pearce on the structure of English from cradle to grave—what can be, should be, first taught, what next, and why. We do have before us the completely integrated curriculum guide produced over several years by summer workshops in the Tamalpais Union High School District in California, and a similarly effective and intelligently devised guide which is the result of the five-year Portland High School Curriculum Study; and there are more of these excellent things, completed or in the making, put together through the combined efforts of experienced teachers, subject matter specialists, and experts in curriculum. These are remarkable achievements, worthy of great applause and rigorous trial.

But need we not also ask ourselves whether such a stalwart champion as George Winchester Stone is really correct in insisting that the study of English in our schools should be cumulative and sequential, levels of instruction building, like blocks, upon other levels until a final structure containing all, or most, things is erected? May we not ask again whether literature can truly be structured this way or any way, and how we can find out about whether it can be? We need to remember Lionel Trilling's warning that though in a certain sense literature provides a discipline, at the same time it abhors a discipline. We need to be reminded again of Northrop Frye's assertion that literature is not a subject of study, but an object of study, and T. S. Eliot's that foreign literatures can be taught because each involves the mastery of a language in which instruction is often necessary, but that to study the literature of one's own language is an absurdity—it needs only to be experienced. We can say that literature cannot be taught, but only ways of talking about literature, or we may finally submit, as many of us do, to the difficulty of doing that and so justify our talking in terms of those disciplines which encroach upon literature—history on the one side, philosophy on the other.

Perhaps all of these things are absurd, for they are expressed by poets who, as Pindar said he did, speak "things like truth" rather than truth itself, so that men like Plutarch could call all poets liars, and Plato would dismiss them from his land. But they speak of areas which, insofar as our concern is literature, how to read or talk about it, are of interest to us all. If I am correct in supposing that the research methods of literary scholarship cannot prove useful in educational research, I think I am correct also in supposing that the results of literary scholarship and of educational scholarship can be mutually helpful and invigorating. That we meet here to talk about these and about other interests which as teachers we have in common is a good thing; but it will be, I think, a lesser thing if each of us does not leave this conference to return with increased vigor and more humble understanding to the cultivation, if you will excuse the expression, of his own vineyards, knowing that we can talk plainly together, but knowing also that, as our great educator and critic, Henry David Thoreau, once said, "We must first succeed alone, that we may enjoy our success together."

RESEARCH METHODS OF THE LINGUIST WHICH MAY BE USEFUL IN EDUCATIONAL RESEARCH

Albert H. Marckwardt

As a first step in assessing the possible contributions of linguistics—I speak here of the discipline rather than of the individual scholar—to research design in education, we must understand clearly what the linguist is trying to do and wherein his aims differ from those who are engaged in educational research. Today, perhaps more than ever before, the linguist views language as patterned human behavior, as a system, or even more appropriately, as a series of interlocking systems. His aim with respect to these is primarily descriptive: to describe each system in a manner which is at once accurate, complete, consistent, and economical.

Such a description may be of a language considered in its totality, whether it be spoken by two hundred or two hundred million. It may be confined to a regional or local dialect of the language. It may be limited to the language as spoken by a single individual, for which the term *idiolect* has been coined.

Linguistic studies are customarily divided into two great classes: synchronic and diachronic. Synchronic investigations have as their aim the description of a linguistic system at a particular point in time. Often it is a language in its current state, but it may be the English language at the time of King Alfred or Latin at the time of Julius Caesar. Diachronic studies compare the structure of a language at two or more points in time. A history of the English language is a diachronic study.

Pattern and Behavior

What has been said in this all too brief introduction points to one very fundamental difference between much of linguistic and educational research. The primary concern of the former is the description of a system, what de Saussure characterized as *langue*. Educational research, on the other hand, focuses its attention principally upon the linguistic acts or possibly more often upon changes in the way in which the system is manipulated as a consequence of certain planned experiences (teaching devices, practices, etc.). This concerns the other member of the de Saussurean dichotomy, namely *parole*. Or, returning to our original characterization of language as patterned human behavior, linguistics is concerned with the pattern though always bearing in mind its behavioral manifestation; educational research fixes upon the behavioral manifestation, although this must necessarily be considered as capable of description only in terms of pattern.

From this initial description it is evident that despite the differences in focus, there is as well an underlying unity in that linguistics and educational research in English are unquestionably dealing with the same phenomenon: the sequences of sounds that human beings employ in communicating with each other and the manner in which these sounds may be represented in writing. It is a reasonable supposition, therefore, that approaches and modes of study which have aided the linguist in dealing with *langue* may also assist the research specialist in education when he concerns himself with *parole*.

Analysis of a Sample

Possibly the first point to be made is that essentially all linguistic study consists of the examination and analysis of a sample. Only rarely is a language studied in its totality. This occurs chiefly when some earlier and now extinct language is available to us only in fragmentary remains, as is the case with Tocharian or the Minoan inscriptions. Even in the case of a still extant language such as Chitimacha with only two speakers remaining, which was described by Morris Swadesh some years ago, it would be idle to pretend that Swadesh dealt with the total linguistic output, either actual or potential, of these two individuals.

Certainly no one could pretend to study a language such as Modern English in its entirety, namely to examine everything that is being said and written by every speaker of the language during any one year, week, or even a day. The mass of material would be so great as to defy handling, nor is it certain that if it were collected and analyzed, any great advantage would accrue.

Dialect research regularly depends upon sampling techniques, some of them fairly complex. The historical linguist, as he probes the language of former eras, also deals with less than the totality of what was written and spoken. In certain instances—Old English is a case in point—only a relatively small and certainly far from representative sample of the language of the age of King Alfred has been preserved. Here we must make the best of what we have. But the English of the Renaissance has come down to us in a quantity so great as to prohibit total analysis. Again, selectivity in what is to be studied is virtually mandatory.

In a number of areas of linguistic study American scholars have established a reputation for both ingenuity and extreme care in the selection of samples. Ingenuity is demonstrated by Charles C. Fries's use of letters sent to a government bureau as the corpus for his *American English Grammar.*[1] These afforded him not only the materials for study but also gave him complete data on the age, place of birth, residence, occupation, and education of the writers; in short, a number of control factors were built in. Another instance of ingenuity is to be found in a recent doctoral dissertation in which the author compared the language of taped panel discussions by certain public figures with articles written by the same persons, in order to determine the use the extent of subordination in spoken and written English. Someone has been taping the *What's My Line* television program in order to obtain a body of unrehearsed oral materials for the study of interrogative intonation patterns.

Representativeness of Sample

So much for ingenuity. The chief demands upon a language sample, however, are that it be representative, sufficient but not too unwieldy in extent, and comparable to other samples which may constitute the entire corpus selected for study. Perhaps no linguistic project has exercised greater care in this respect than the various regional investigations which, taken together, comprise the Linguistic Atlas of the United States and Canada. Here the problem is to extract from a limited number of informants a body of data which reflect the extremely recessive dialect features of American English. Because of this very recessiveness, random sampling would not suit the purpose. To avoid the admixture of speech forms from other areas, only persons who were born in the community in question and have lived in it virtually all of their lives are selected for the interviews. Sometimes they are of third and fourth generation families in the particular area. To reflect the influence of constantly more extended schooling over the years, every community selected for study is represented by two informants, one with little more than a grade school education who is seventy years of age or older; a second with some secondary school education who is between forty-five and sixty. In certain cities, particularly those which are known to be focal points for the spread of speech forms, cultured informants, namely persons with some exposure to college or those whose principal social contacts are with such a group, are also interviewed. This is what might appropriately be called stratification sampling; it is a conscious practice devised to meet a situation where random sampling would reveal very little.

I am aware, of course, that educational research also leans heavily for its validity and effectiveness upon the use of sampling techniques. However, the educational researcher is usually concerned with sampling the population, with developing rigorous controls in the selection of student groups, rather than with sampling of the language. Nevertheless, it is conceivable that overlooking the criteria of representativeness and comparability in the language samples under scrutiny might also weaken the design of a research project in education.

---

[1] Charles C. Fries, American English Grammar (New York: Appleton-Century-Crofts, 1941).

Limitations of Samples

This has been particularly apparent in some of the recent attempts to study the spoken language of elementary and secondary school children. The danger here is two-fold: extracting the speech and writing samples from communication situations which are not comparable, and drawing conclusions which go beyond what the sample materials will justify. For example, to base an analysis of the structure of children's speech wholly or principally upon classroom discussions in which the students are struggling to formulate and express ideas far more complex and profound than those which they normally communicate is certainly calculated to give an impression of incoherence, of mixed syntax, of false starts, of frequent hesitation pauses. Although this may fairly reflect one type of linguistic performance, it is by no means typical of the totality of children's speech. Consequently, due caution must be exercised in whatever conclusions are to be drawn from such a sample.

The danger of extrapolating too much from a limited sample is illustrated by a study to which I referred somewhat earlier, one which was concerned with the amount and nature of clause subordination in English. The comparison of the spoken and written sentences of a number of political commentators and columnists resulted in the surprising discovery that these individuals used a higher proportion of complex sentences in their speech than in their writing, a conclusion running directly counter to what has generally been assumed to be the case. Unquestionably the conclusion was valid for these two bodies of material; its extension to the English language generally would be dubious indeed, simply because most of us do not do most of our talking in panel discussions concerned with complex topics.

Some thought must also be given to the connection between the nature of the speech situation and the particular grammatical forms or syntactic structures in which the investigator is interested. It would be inefficient, for example, to employ material which is primarily narrative for a study of present tense verb forms and syntax, just as it would be unproductive for a student of French or German to use essays as a corpus for the analysis of the second personal pronoun.

Another linguistic area which is frequently involved in educational research design and about which there is considerable confusion is that of vocabulary frequency and word counting. The confusion arises partly from the method of sampling and in part from vagueness of definition. It was amusing to observe, some years back, that the dramatic increase in estimates of the size of individual vocabularies coincided neatly with a shift from abridged dictionaries containing approximately 125,000 entries to the unabridged with a content of half a million as the basis for the sampling.

Moreover, it is often difficult to determine what is being counted or indeed what should be counted as a word. The question confronts us at both ends of the lexical scale; we may as well begin with the upper end, which involves the compounds. Should *school board* be counted as a word, apart from its two component elements? Possibly so, since its meaning might not be comprehended simply from the combined meanings of the two words which are its component parts. If we count *school board,* would the same argument prevail for *schoolroom?* A *schoolroom* is patently a room in a school, and although it takes on certain unique characteristics, the meaning is virtually self-evident. The latter is even more true of combinations like *schoolboy, schoolgirl,* and *schoolteacher.* If these latter are to be considered as individual lexical items, then almost any two-word combination can reasonably be so counted. Let me hasten to say that the manner of writing these words, whether solid, hyphenated, or as two, cannot be made to serve as a guide; the practice of the language is notoriously inconsistent in this respect.

A similar problem exists with the addition of derivative suffixes. For example, *-ness* can be affixed to almost any adjective to form an abstract noun; *-er* can be added to all verb stems except the modal auxiliaries to produce an agentive noun. Here are two features of the language which constitute working parts of the linguistic equipment of everyone for whom English is a first or native language. Whether a particular combination of a root with *-ness* or *-er* has been recorded or not is quite immaterial; it is potentially there. But taking these two features of the morphological structure of the language into account would, in effect, create a series of nouns equal in number to the combined total of virtually all the verbs and adjectives. This would have important implications for any fragmentary count which was being conducted.

Prefixes pose something of the same problem. *Anti-* can be added to almost anything, and once an opposition attitude has thus acquired a name, the counteropposition responds with a new combination employing the prefix *pro-*. Allen Walker Read once commented that the history of opposition movements in the United States could be gleaned from a study of the combinations with *anti-* appearing in the *Dictionary of*

American English, and of course the same is no less true of the English language as a whole, though on a much larger scale.

Inflections also pose a series of knotty questions although the numerical implications may not be quite so great. Are noun plurals to be counted as items distinct from noun singulars? If regular plurals (*cat, cats*) are not to be so counted, what about the irregulars? Are *mouse* and *mice* two words or one? If these should be counted as separate words, possibly on the ground that the stem vowel differs, would the same principle apply to *calf, calves,* which involves consonant rather that vowel alteration? And what about such plurals of specification as *filings, sweepings,* which are quite distinct in meaning from the gerundives which form the basis for the singular: *filing, sweeping*?

Similar questions arise with other parts of speech. If *good, better, best,* positive, comparative, and superlative respectively of an irregular adjective, constitute three separate words, how about *tall, taller, tallest?* Would the decision on *tall* differ from that on *long,* in which the *-ng* combination is pronounced differently in the comparative and superlative from what it is in the positive? Irregular and regular verbs pose precisely the same sort of problem, witness *sing, sang, sung,* as compared with *love, loved.* A similar series of questions may be asked with respect to the subject and object forms of the personal pronouns.

Nor can we overlook the problems posed by functional change. Is the noun *hate* a word distinct from the verb *hate?* How about a word such as *down,* which may be used in no less than five different part-of-speech functions? Is the gerund, ending in *-ing* but used as a noun, the same word as the participial form in *-ing* used as an adjective?

Semantic counts are possibly even more perplexing. Often a semantic count for English will employ quite arbitrarily the major division of word definitions as given in the *Oxford English Dictionary* as the unit for counting. But the *Oxford English Dictionary* in the course of its fifty-year editorial process had at least four editors-in-chief. Some of them dealt with word meanings by recognizing relatively few major divisions but many subheadings or minor divisions. Others worked in exactly the opposite way. In the light of this wide variation in editorial procedure, how much confidence can we have in what is being counted as a unit?

My purpose in raising these questions is neither to befuddle nor to mystify. Nevertheless, it is quite evident that two separate counts of the same body of material might result in quite different conclusions as to the total wordage and consequently affect estimates of frequency as well. Actually, one concordance of Shakespeare's plays contains 24,000 entries; another has just 15,000. These two counts are based upon precisely the same body of material. Nor do these two numbers by any means constitute the conceivable limits of variation. The important thing in matters of this kind is to know what constitutes a unit. No word or frequency count can be accepted seriously unless we know exactly and unequivocally how the counting was done.

Related difficulties arise in connection with measures of the frequency of error, also a feature common to many research projects in education. Several problems arise in connection with these. The most obvious, of course, is the question of what counts as an error. Here a keen eye to the realities of standard usage and the ability to employ and interpret scholarly sources of information about the language is the first requisite. A further question is the relative weight to be given to types and tokens. A final problem is how broad a range may be included within a single error type. Again unmistakable clarity as to what is being counted and how the counting was done is an absolute necessity with respect to any linguistic unit whether it be a phoneme, morpheme, or syntactical item.

Primacy of Spoken Language

One approach of the linguist to the material with which he is concerned that is most important for educational research is his attitude toward speech and writing respectively. The linguist considers the spoken language as a primary, the written language as a secondary or derived manifestation or form of linguistic behavior. There are many reasons for this. Limitations of time and space prohibit a full explanation.

It is enough to say, perhaps, that historically considered, man has been speaking for at least 100,000 years, possibly more. He has been writing for no more than 6,000. Each of us learns to speak some years before he learns to write, and by the time that he gains an elementary mastery of reading and writing, he has already acquired a considerable control of the language. Of the more than three thousand languages of the world, there are hundreds that are spoken but have not been reduced to writing. The converse is not true: There are no written languages which are not or have not been spoken. Even in this highly literate age, every one of us speaks far more than he writes. Our weekly spoken output has been estimated as approximately equalling a short novel. Obviously, we write much less than that.

More important than any of these considerations is the fact that the patterned structure is often more immediately apparent in the spoken language. For one thing, the written language is forced to employ various devices to compensate for its inability to portray stress, information, and juncture, all of which function grammatically as well as rhetorically.

2

This is not to say that the linguist has no interest in or concern with the written language. If he hasn't, he should have, particularly if he is dealing with English or any other language with a long written tradition. There is no doubt, however, that to consider written English in terms of its differences and deviations from the spoken language is more realistic and more productive than to approach the matter from the opposite direction.

In this connection it is also important to clarify the meaning of the term *colloquial* as it is used by linguists and lexicographers. It does not refer to an inferior or a degraded kind of language; it means merely language in its spoken as distinguished from its written form. In this sense it indicates a functional variety of language and not a social level. The frequently employed diagram which places literary or written English at the top of a hierarchical arrangement, colloquial English below it, and popular or vulgar English at the bottom is confusing functional variety with social level. It is most important that muddy thinking with respect to this should be avoided in current educational research.

Reading Interprets Writing System

We must not fail to recognize, moreover, that when we begin to teach reading in the lower elementary grades, we are by no stretch of the imagination teaching language. The six-year-old, upon his entrance into school, already has a considerable command of his linguistic medium. Except for slight individual and incidental difficulties, he has mastered the sound system, including intonation and stress features as well as the articulation of the individual phones. He handles the regular inflections competently within the limits of his vocabulary; irregular noun plurals and verb tense forms may cause occasional trouble. He employs the basic word order patterns in their simple, if not in their expanded forms. In short, his linguistic sophistication is advanced far beyond the dreary iteration and two-word sentences which portray the rollicking adventures of Dick and Jane. It is not possibly until he reaches the fourth grade that reading will become for him a means of extending his command of the language. Up to that time, learning to read is by no means synonymous with learning language; it is training in the manipulation and interpretation of the writing system. This has not always been sufficiently recognized in designs for educational research.

Another factor of which the linguist is usually very conscious is the totality of the linguistic context for any individual. In the daily linguistic experience of, say, the child of twelve, the interplay of language which goes on within the school during the five hours of formal instruction is by no means as extensive or as natural as the communication which takes place in the eight or nine hours which he spends on the playground and in the home. Moreover, two quite disparate dialects may be represented here. The direct teaching of English to which he is exposed constitutes an even smaller proportion. To couch educational research in terms of designs which neglect or underrate the impact of ordinary day-to-day linguistic interchange and which ignore the patent artificiality of much of the schoolroom context is certainly quite unrealistic.

Nor can we afford to overlook, in our estimates and attempted evaluations of various instructional methods and devices, the socially determined attitudes of the child toward the language he encounters in the school: toward that which is employed by the total group of teachers and that which is recommended by his instructors in English. I hasten to point out that the two are not necessarily and always the same.

On a certain socioeconomic level, the pupil may be totally indifferent or even feel a degree of hostility toward a social dialect which is not his. Naturally, this is not always the case. In some situations the classroom teacher, representing the principal cultured or quasi-cultured contact which the child encounters, may exercise a linguistic influence of considerable proportions. But we must also be aware of the possibility of an attitude such as that expressed by Nancy Mitford, commenting on the situation in England: "Any natural dialect that is still left to the children will have superimposed upon it the language of the primary school teacher (a class of people entirely non-U[pper class]) so that children leave school speaking a mixture which has nothing to recommend it." I am not here concerned with the justice or accuracy of Miss Mitford's remarks. I cite them merely to point out that language attitudes have a great deal to do with the effect of instruction upon language behavior.

Language Is Behavior

The final point I shall make is centered about the behavioral concept. It is not unusual these days for educators to conceive of or describe learning as change in behavior. The linguist, as we have already seen, is prone to characterize language as patterned human behavior. Consequently, when he proceeds to analyze a language or to describe its structure, he feels it necessary to do so in terms of formal manifestations of that behavior rather that in terms of ideas or logical concepts. The following paragraph from a recent book by Hall of Cornell University makes this point very succinctly:

It has taken centuries for students of language to develop a scienti-

> fic approach to their subject matter and a corresponding analytical technique. In order to attain these they have had to strive for the greatest possible objectivity, viewing language primarily as a means of communication among the members of a society, and only secondarily as a medium for the expression of individual personalities. When describing linguistic structure, its analysts have aimed at establishing functional units (phoneme, morpheme, and the like) in terms of which they could state as exhaustively as possible the relations among elements of language.[2]

What is of importance here for the research scholar in education is perhaps not so much a knowledge of the actual techniques and terminology of the linguist as an understanding of some of the basic assumptions from which he proceeds and the attitudes toward language and languages which he maintains. First of all, there is the same attempt to maintain objectivity in dealing with linguistic behavior that social scientists strive for in other fields. This is achieved in part by recognizing functional units as the basis for analyzing linguistic structure, and identifying these functional units in terms of formal criteria: inflections, function words, word order, and the like. Thus, the linguist in dealing with the area of verb structure conventionally labeled the *subjunctive* would begin, not with such mental or ideational concepts as contrary to fact, hypothetical tentativeness. He would instead fix his attention first upon the use of verb inflections and various types of auxiliaries, considering their appearance and distribution in relation perhaps to the types of words introducing the clauses in which they occur. Only after having made such a formal inventory would he feel it appropriate to work back to the ideas involved.

In considering language as a socially and culturally determined activity, the linguist recognizes that geography, social level, and stylistic mode all affect the structure of the linguistic system in different ways, placing certain limits upon the extent to which the individual can alter or manipulate the medium. Indeed, it is no mere coincidence that within the last two decades the concept of structure has become one of major importance both in the science of language and in the philosophy of education.

---

[2] Robert A. Hall, Jr., Idealism in Romance Linguistics (Ithaca, New York: Cornell University Press, 1963), p. 1.

PROBLEMS AND POSSIBILITIES OF RESEARCH IN THE TEACHING OF WRITTEN COMPOSITION

Paul B. Diederich

Let me begin with what I regard as a satisfactory minimal experiment in the teaching of writing. Most of the procedures I describe will also apply to an experiment in the teaching of literature, since the tests used in such experiments will usually be essay tests. Later I shall add a few notes on the special problems and possibilities of objective tests, but the bulk of this paper will be devoted to the problem of making essay evidence sufficiently accurate to be used as a tool of research.

Minimal Experiment

I must preface my remarks by saying that my standards for a satisfactory minimal experiment are quite modest. Some psychometricians give the impression that no conclusions can be drawn from an experiment unless about 200 teachers have participated in a balanced design that provides for every possible contingency. That sort of thing calls for a grant of at least half a million dollars, and I would not sign my name to a request for such a grant unless we had turned up highly significant results in three or four pilot projects of the sort that I shall describe. On the other hand, my standards may seem unduly severe to a teacher who has tried one procedure in one of his classes, another procedure in another, has asked both classes for an essay on the same topic at the end, has gathered the impression that one class did better than the other, and then wants a large grant from the Ford Foundation to prove that his impression is a universal law. Such do-it-yourself procedures hardly warrant a request for a grant.

I have used, in various studies, all of the procedures I shall describe, but here I shall combine them in an imaginary study, leaving out the features that proved unworkable or unnecessary. It will also facilitate my exposition if I leave out the accidents, the mistakes, the compromises, and the delays that are inevitable concomitants of research in a real life setting.

This imaginary study was directed by Mr. Archer, supervisor of English, assisted by Mr. Baker, research director, who knew how to compute correlations and the significance of differences between experimental and control groups. Since the people these men had to convince were primarily their board of education, the administrators, the teachers, and the parents, they favored procedures and methods of analysis that they could explain to their audience. They would not settle for sloppy work, but if a highly sophisticated design gained about 2 percent in accuracy and lost 100 percent in comprehensibility, they preferred a simpler, classic design that their audience could understand.

The participating teachers were Mrs. Connor and Mr. Daniel in East High School and Mrs. Ellis and Mr. Franklin in West High School. Each of these teachers offered two regular classes in tenth grade English for the experiment, and Mr. Baker, the research director, found that there was no significant difference between the two classes taught by each teacher in averages or standard deviations on three objective tests: vocabulary, reading comprehension, and ability to detect or correct common errors in writing. I did not say that all eight classes were equal in these respects; I said that the two classes taught by each teacher were approximately equal to one another, or were made equal by transferring one or two students. Mr. Archer, the supervisor, assigned one of these classes to the experimental treatment and the other to the control treatment by flipping a coin.

Thus there were four experimental and four control classes, and each teacher taught one experimental and one control class. The other classes taught by these teachers continued as usual. Mr. Archer released these teachers from one extraclass duty per day but could not offer any other compensation for participation in the experiment.

There were six other English teachers in each high school, and each of the sixteen teachers was assisted by a reader. Two or three of these other teachers in each school also volunteered for the experiment but were rejected because their regular tenth grade English classes were too dissimilar in verbal ability. (There may have been some additional reasons in a few cases, but these reasons were not mentioned.)

The experimental and control treatments can be anything you like, but I have chosen two to illustrate how one thing leads to another. Mr. Archer had been trying for some time to convince his superintendent and board that teachers assisted by readers get better results than teachers unassisted by readers. He had finally begun to find significant differences in the expected direction, but these differences have been smaller than he had anticipated. To find out what was wrong, he had analyzed 320 papers—ten drawn at random from the files of each teacher and reader—and found that more than 80 percent of the annotations of both groups were corrections of specific errors in usage, punctuation, spelling, and wording. He then had given some inventories provided by

Educational Testing Service and found out that these errors were being eliminated at the average rate of only 2 percent per year—somewhat less than could be accounted for by dropout if the teachers had done nothing at all. This finding had led him to the hypothesis that both the teachers and the readers were doing the wrong thing to the papers, and hence that it made little difference in the quality of student writing whether the wrong thing was done by one person or by two.

Casting about for two sharply contrasting methods of treating papers that would show whether even extreme differences in treatment made any difference in results, he had hit upon the notion of using *positive reinforcement* in the experimental classes and *comprehensive correction* in the control. In the first, the teachers would examine each paper paragraph by paragraph for anything the student had done well, or even a little better than before, and would show him that these points had been noted and appreciated. They would also limit themselves to not more than one suggestion for improving the next paper. In the control classes, these same teachers would mark every error that they could find. They would not avoid praise altogether, but they would use it sparingly.

When these treatments were discussed in staff meetings, the objection was raised that it would sound hypocritical if twelve good things were noted in a paper, only one suggestion for improvement was made, and then the paper was marked D. It was finally agreed that in both experimental and control classes, the instructional (homework) papers would not be graded at all; the teacher's reaction would be conveyed only by comments and corrections. It was decided that this could be done without throwing the students into the tizzy that usually attends the withholding of grades, because these teachers had already agreed upon an unusually rigorous method of evaluation for all their future experiments in the teaching of writing. All classes involved in these experiments would write a short test paper in class on or about the first of each month from September to June, making ten test papers in all, the ratings of which would be revealed to the students.

Not even the teachers would know the topic for any test paper before the day on which it was to be written. Instead, they agreed in advance upon a long, numbered list of topics acceptable to all as topics for short, impromptu compositions. Thereafter, at the opening of each testing day, Mr. Archer announced, "Today we'll use 63," or "Today we'll use 41." It was not thought necessary to pick one topic for morning classes and another for afternoon classes, because the teachers had found no significant differences between morning and afternoon grades on the same topic. They only wanted to avoid the suspicion that, if the topics were known in advance, some students would write a paper at home, have it edited by their parents, and practically reproduce it the next day, while others would not have this advantage. They could not prove or disprove that such assistance made any difference, since no one would own up to it, but they knew that their findings would be more convincing if this possibility were eliminated.

The test papers were all written in standard four-page folders provided by the school, with lines spaced half an inch apart for easy insertion of corrections by the students. Nothing was written on any test paper by any teacher or reader lest it bias the judgment of some future reader. Their ratings and comments were all made on separate cards.

The test papers were numbered by the students themselves with any number of six digits that popped into their heads, such as 924336 or 004395. Six boxes were printed in the top right-hand corner of each folder with the directions, "Write any number from 0 to 9 in each box." Duplicate numbers were reduced by adding, "Avoid numbers like 555555, 123456, 987654." No other identification was written on any folder. Each student copied his number on a 4 x 6 *record card* like the following:

Grade .... Curriculum .... Paper ....
Teacher ....
Student .... Year ....

	Sept.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	June
SUM 1										
SUM 2										
TOTAL										
Prev. Total	X									
Cum. Total	X									

These record cards were turned over to Mr. Archer at the end of each testing day and returned to the four experimental teachers only after two independent ratings of each test paper had been turned in: one by a teacher other than one of the experimental teachers, the other by one of the readers. Neither of these had any direct contact with an experimental or control class except for these ratings. Teachers of the experimental and control groups had nothing to do with the rating of test papers; that was done exclusively by other members of the department and their readers.

Each judge recorded his ratings on a 3 x 5 *rating card* like the following:

Topic .... Reader .... Paper ....
1–Weak 2–Fair 3–Good 4–Better 5–Best
Quality and development of ideas 1 2 3 4 5

Organization, relevance, movement	1	2	3	4	5
Style, flavor, individuality	1	2	3	4	5
Wording and phrasing	1	2	3	4	5
Grammar, sentence structure	1	2	3	4	5
Punctuation, capitals	1	2	3	4	5
Spelling	1	2	3	4	5
Handwriting, neatness	1	2	3	4	5
Sum of ratings	__________				

Each judge encircled the number corresponding to his rating on each quality. (The empirical evidence behind these headings will be discussed later.) On the last line, he recorded the sum of the numbers he had encircled. This number was checked by each student before he recorded the sum on his record card, because there could be many mistakes in addition. If a student found such an error, he did not correct it himself but took it to his teacher, who recorded the correct sum with a red pencil. The five numbers after each quality could be thought of as grades of F, D, C, B, and A, but they were written as numbers to facilitate addition. (At some future time, it may be found desirable to give some of these ratings more weight than others, but present evidence from factorial studies gives no firm basis for introducing this complication.)

While the papers were being rated, the record cards were locked up—which forestalled any suspicion that the raters might learn the identity or previous history of the writers from the files of record cards. Afterwards, they were returned to the four experimental teachers so that each student could copy the sums of the two ratings on his record card, add them to get the total rating on each test paper, bring down the previous total (after the first paper), and add it to the present total to get the cumulative total down to that point. His copying and adding were checked by some other student. These cards remained in the possession of the experimental teachers until the close of the next testing day. In each test, if any student forgot his number, his teacher referred to the cards and informed him. This procedure had proved simpler than choosing a new number for each test paper, which made the compiling of records infernally complicated.

After all the papers had been collected on each testing day, an assistant arranged them in the order of these self-chosen numbers, which put them into a completely random order. Then one pile of approximately thirty papers was given to one of the cooperating teachers to rate, and another pile of the same size was given to one of the cooperating readers. After each had completed his rating, he returned the papers and rating cards to Mr. Archer, who kept the cards but passed the papers on for a second rating by another teacher or reader. The papers first rated by a teacher were rated the second time by a reader, and vice versa. The second rater was not told who had rated the papers previously and had no access to the previous ratings.

In an experiment like this concerning writing, it was important to judge *writing*—not the character of the writer, which would inevitably color the rating of anyone who knew the student and recognized his handwriting. Since it was known that raters have biases in some directions and blind spots in others, it was also important *not* to have the same judges rate the same papers every month. Papers with the thirty lowest numbers were given to the first judge the first month, and papers with the next lowest numbers were given to the second judge. The next month, these papers were given to the third and fourth judges, while the first and second took up where they left off the month before. The aim was to have the papers of each student rated by all cooperating teachers and readers in the course of the year. There was no undue burden on any one person, since he had to rate only about sixty papers a month: thirty as a first rater and another thirty as a second rater. (In large departments, it is possible to reduce this burden to thirty papers a month, but while the rating procedure is being learned, it is usually thought advisable for each rater to have more practice than this with the use of the rating scale. If it is possible, it is also desirable to pretest the raters by having them all rate twelve to twenty sample papers in a one-day workshop just before school opens in the fall. They can then discuss the reasons for their differences of opinion and agree upon definitions of high, middle, and low points on each quality in the rating scale. Teachers and readers whose judgments are extremely erratic, as judged by their deviations from the consensus on each paper, can be given special duties other than the rating of test papers.)

Significance of Differences

Reference to the above rating card shows that if a student was "good" in every respect, he received a total rating of 24 from the first judge and 24 from the second, making a grand total of 48 on each test paper. If he continued at that level throughout the year, he obtained a cumulative total of 480 points on the ten test papers for the year. Previous experience suggested that the "standard deviation" (the spread of scores around this average) would be about 100 points. This figure was computed for each group, experimental and control, by Mr. Baker, the research director. To find out whether there was a "significant difference" between the average cumulative totals of the experimental and control groups at the end of the year, he used the following procedure:

1. Find the "standard error" of each average. This is the standard deviation of each group divided by the square root of the number of students. The standard deviation, let us assume, is 100 points. There are four classes in each group, making about 120 students, and the square root of 120 is approximately 11. 100 divided by 11 is about 9, the standard error of each average. The formula is:

$$S.E._{av.} = \frac{S.D.}{\sqrt{N}}$$

2. Square the standard error of each average. 9 × 9 = 81.
3. Add the two squares. 81 + 81 = 162.
4. Take the square root of this sum. The square root of 162 is 12.73. This is the standard error of the *difference* between the two averages.
5. All the figuring down to this point may be represented by the formula:

$$S.E._{diff.} = \sqrt{S.E._{A}^{2} + S.E._{B}^{2}}$$

in which S.E. means "standard error." A = experimental; B = control.

6. Let us imagine that the experimental group (positive reinforcement) came out 30 points higher in average cumulative totals than the control group. There are now two crucial numbers to learn by heart. The first is 2. If the difference between two averages is *twice* the standard amount of that difference, the difference is "significant at the .05 level." Twice the standard error of this difference (2 x 12.73) is 25.46. The difference between the two averages (30 points) is therefore significant at the .05 level, since it is larger than 25.46. A difference this large will occur less than five times in a hundred trials under exactly the same conditions if there is no true difference. Roughly speaking, therefore, we are at least "95% sure" that this difference is "real"—not caused by random fluctuations in our measure.

The other crucial number to learn by heart is 2.6. If a difference between two averages is 2.6 times as large as the standard error of that difference, the difference is "significant at the .01 level." A difference that large will occur less than once in a hundred trials under exactly the same conditions if there is no true difference. Then we are "99% sure" that the difference is "real"—not caused by random fluctuations in our measure. In this case, 2.6 × 12.73 = 33.098. Since this is larger than the obtained difference of 30 points between the two averages, we did not reach the .01 level of significance. However, the chances are still 19 out of 20 that the ob-

> tained difference was "real"—not in the chance domain. This is quite large enough for a practical judgment that the method of *positive reinforcement* was—in this situation and by these measures —superior to the method of *comprehensive correction.*

Nearly all conclusions of quantitative experiments in education are stated in these terms. The differences found are usually tagged with one asterisk if they are significant at the .05 level, with two asterisks if they are significant at the .01 level, or with "NS"—not significant—if the difference does not reach even the .05 level of significance.

There are three things to remember about these ground rules of educational research. First, the notation NS—not significant—does not prove that there was no true difference between the two things that were being compared; it only indicates that no true difference was proved. In general, we do not want to risk large sums of money in education on differences that might have come about by chance. Second, if a difference is found to be significant, this does not necessarily mean that it is important. When whole states are tested, differences as small as a hundreth of a grade-point may turn out to be significant, but they have no practical consequences.

Most important of all, in view of the conclusions drawn from some current research, is the fact that if no significant difference is found between two treatments, and if highly valid and reliable tests were used, then there is probably nothing to choose between them. One is about as good as the other. But if the tests were not appropriate, and particularly if they were not sufficiently reliable to detect the relatively small differences that one might reasonably expect, then they prove nothing whatever. I am particularly wary of experiments in English composition when the reliability of grading the test essays may be assumed to be at the usual level of about .5. This reliability is seldom reported, but one can estimate it from the reliability of grading in other experiments in which similar grading procedures were used. With this level of reliability, one could hardly expect to find a significant difference between two treatments of any sort: much writing vs. none; new grammar, old grammar, or no grammar; or even taking English vs. not taking English. The conclusions, however, are sometimes written up in this fashion: "Since no significant difference was found between much writing and little writing, we must conclude that the amount of writing makes no difference, and we must now turn our attention to factors that matter"—i.e., those that have not yet been tested.

Nonsense! There may have been a large and important difference, but the measures were not sharp enough to detect it. Whenever I see that

NS signal, I look at the reliability of the measures, or estimate it if it is not reported; if it is not at or near .9, I conclude that the researcher did not find any difference, but that with such wobbly measures he could hardly expect to. With a tape measure printed on elastic, I could not find a significant difference between the height of my wife and that of my grandchild, but that would not prove that no such difference exists.

I honestly believe that almost all experiments concerning English composition that rely on essay grades have been conducted with tape measures printed on elastic, and that they must all be replicated with measures in which we can have confidence. Even the long-standing conclusion that knowledge of any sort of grammar correlates with writing ability about .3 is probable—in view of the number of investigators who have come up with approximately this figure—but not, to my way of thinking, clearly proved. With the measures of writing ability that were used, one could hardly expect to find a correlation higher than .3 with anything.

I recommend the following as a standard method of computing the reliability of grades or ratings on test papers. Have two papers written on topics requiring the same mode of writing, preferably on different days. Have paper 1 rated by Judges A and B; have paper 2 rated by judges C and D. Then correlate the sum of the two ratings on the first paper with the sum of the two ratings on the second paper. This procedure takes into account both variations in judges and variations in students from one paper to another. With ordinary grading procedures, the best you can expect is that the correlation between the two sums will be about .5. This is the reliability of one paper, graded in the usual fashion by two judges. To find out how many such papers you need from each student to work up to a reliability of .9, use this formula:

$$\frac{(\text{the reliability you want}) \times (1 - \text{the reliability you got})}{(\text{the reliability you got} \times)\ (1 - \text{the reliability you want})}$$

If you want .9 and got .5 on the first trial, this becomes:

$$\frac{.9 \times (1 - .5)}{.5 \times (1 - .9)} = \frac{.9 \times .5}{.5 \times .1} = \frac{.45}{.05} = 9 \text{ (papers)}$$

Thus, you will need nine such papers, each graded in this fashion by two judges, to work up to a reliability of .9.

Some of the improved rating procedures on which we are now working apparently yield an initial reliability of about .7 for one paper rated by two judges. The formula then becomes:

$$\frac{.9 \times (1 - .7)}{.7 \times (1 - .9)} = \frac{.9 \times .3}{.7 \times .1} = \frac{.27}{.07} = 4 \text{ (papers, approximately)}$$

If the initial reliability of such ratings can be maintained, therefore, we shall need only four such papers from each student, each paper rated in this fashion by two judges, to work up to a reliability of .9.

For the time being, however, we are insisting on our monthly test paper, rated independently by two judges, to be on the safe side. One reason is that the new rating procedure is still in the early stages of tryout, and there may be some artifact, not yet apparent, that spuriously boosts reliability. Second, an experimental result is seldom maintained at the same level in daily classroom practice. Third, we have to make allowance for human error. While teachers are learning the new system, some may make large-scale errors that will invalidate whole sets of papers. For example, whole sets of record cards may be mislaid or ascribed to the wrong students. If one secures about twice as many measures as are theoretically necessary, such losses need not be disastrous.

Some Problems in the Comparison of Gains

Most research designs insist on having a direct measure of the ability in question at the beginning of the experiment and a parallel measure of the same ability at the end. Then they find the difference between the initial average and the final average for the experimental group and for the control group and call this difference their *gains*. The procedures outlined above are then applied to find out whether there is a significant difference between the *gains* made by the two groups during the experiment.

In the case of writing, this procedure involves getting an initial essay and a final essay on the same choice of topics from both groups and having both sets of essays mixed together without identification and graded at the same time—probably during the following summer—by readers who do not know which essays are initial, final, experimental, or control. Then one subtracts each student's initial essay grade from his final essay grade and calls the difference his "gain." One finds the average and standard deviation of these gains for the experimental group and for the control group and follows the procedures outlined above to find out whether there is a significant difference between their gains.

I myself have recommended this procedure both in ETS and NCTE publications. I wish now to retract this recommendation and to say that I would not touch this procedure again with a ten-foot pole. It used to work for me when I was dealing with a mature and stable college composition staff who had been working together on the grading of examinations for several years. Since that time, I have tried this procedure repeatedly in field studies involving better-than-average high schools, and

I have never yet succeeded in getting a significant difference out of it, no matter how striking were the differences in experimental and control treatments, and no matter how many students were tested.

The basic difficulty is that the reliability of a single essay as a measure of the starting point or end point is extremely low, and the reliability of the *difference* between two such essays—that is, the reliability of the gain—is so much lower than the reliability of either essay that it is almost never significantly different from zero. Then one can use 3,000 experimental students and 3,000 control students (as I did in one experiment) and find no significant difference between their gains—simply because the reliability of each gain is practically zero, and 6,000 times zero is still zero, as I should have learned in the fourth grade.

What makes this procedure almost suicidal is the formula for the reliability of a difference, or a gain. One takes the average reliability of the initial and final essays and subtracts the correlation between them. Divide this number by one minus the same correlation. This is represented by the formula:

$$\text{Reliability diff.} = \frac{\text{Av. reliability i,f.} - \text{correlation i,f}}{1 \qquad \text{correlation i,f}}$$

(The designations "i" and "f" stand for "initial" and "final" essays.)

When one takes account of variations in judges and variations in students from one essay to another, the reliability of a single essay in field studies is usually between .3 and .4. Then, since students who write good essays in the beginning tend to write good essays at the end, the correlation between initial and final essay grades is almost as high as the reliability of either grade. Then we get something like the following:

$$\frac{.3 - .2}{1 - .2} = \frac{.1}{.8} = .125 \text{ (reliability of gain)}$$

Since that reliability is not significantly different from zero, it practically guarantees that no significant difference will be found, even if the best English program in the world is pitted against the worst. That is why I do not rely on single essays as measures of beginning and end points, and why I will not let myself be maneuvered into comparing gains from one point to another if I can possibly avoid it. I now advocate a return to something like the method of matched groups, but with a difference. Each student in the experimental group need not be matched with some student in the control group in every possible respect. Instead, there should be no significant difference between the experimental and control groups in initial averages and standard deviations in question—in this case, writing ability. I am even so bold as to advocate the use of objective tests for this purpose of initial equating of experimental and control groups. I myself would use just three tests: reading comprehension, vocabulary, and ability to detect common errors in standard written English. These three tests correspond roughly to the College Board Scholastic Aptitude Verbal Test and its English Composition Test. In combination, these tests have stood up in study after study as the best indicators of writing skill and possibility of improvement at any given time. It may seem odd to use a reading test as part of an index of writing skill, but our SAT Verbal Test is a combination of reading comprehension with a factor adequately represented by vocabulary tests, and it regularly yields higher correlations with good teachers' ratings of the writing ability of their students than most objective tests of writing ability, and far higher correlations than an essay. It is hard to believe that skill in reading and writing can be so intimately related to one another until we recall that our good readers are generally our good writers, and vice versa. Everyone can think of a few exceptions, but it is the general rule that produces high correlations.

It would also be desirable, from the standpoint of convincing teachers and the public, to equate the experimental and control groups initially on the basis of reliable grades on essays, but how can this be done? It takes the better part of a year to work up to a reliable cumulative total of essay grades, using test papers written in class once a month, each graded independently by two readers who do not know the student. In most of the experiments that we are talking about, we get the go-ahead signal in May and have to start in September. Even if we were willing and able to wait a year to secure this initial measure before beginning the experiment, the students so rated would scatter into so many different classes the following year that it would be impossible to assemble experimental and control groups. They cannot be kept together by administrative fiat. Their electives will differ the following year; therefore their schedules will differ; therefore it is impossible for them to remain together in the same English classes. A quicker method must be found to locate groups that are approximately equal in writing ability and possibility of improvement at the beginning of an experimental year. For this purpose, I believe that well-selected objective tests are not merely the only feasible method but are also a reasonably satisfactory method.

For ten years I sectioned classes in freshman composition in a good college on the basis of such tests as these. With the inevitable number of exceptions, the teachers were satisfied that their classes were nearly enough alike in writing ability and talent to make an experiment possible, provided that several sections were used, and provided that the sections matched with one another were assigned to experimental and control treatments by chance. This prevented systematic bias in favor of either group, and minor differences tended to cancel out. We had also tried to section on the basis of an initial essay written during orientation week, but we found that these essays were such atypical performances and the staff disagreed so wildly as to their merits that the objective tests were a safer bet. This is borne out by the common experience of a great many colleges that have faced the problem of sectioning with an open mind. If we can section on the basis of objective tests, we can also use such tests to certify that the experimental and control classes matched with one another are approximately equal in initial writing ability and possibility of improvement.

Once we have them certified as initially equal, we can simply count the number of points that each group accumulates on its rating cards from month to month and find out at the end whether one cumulative total is significantly higher than the other. Thus we avoid the whole perplexing problem of gains. We assume that both groups were equal in the beginning and we compare the final status of one group with the final status of the other. The reliability of the final status of each group will probably reach or exceed .9. The reliability of the difference between them will not be obscured by having to subtract correlations, because the ratings accumulated by one group will be completely uncorrelated with the ratings accumulated by the other. A further advantage is that we shall not be dealing with small, unimpressive numbers, such as the difference between a final grade-point average of 2.9 for one group and 3.1 for the other. The average cumulative total for one group is more likely to be on the order of 470 points as against 500 points for the other.

I realize that this simple-minded solution of the problem of the comparison of gains will be viewed with the same discomfort by psychometricians that was aroused by Alexander's solution of the problem of the Gordian knot. As you remember, he did not untie it; he cut it. At the same time, the psychometricians will welcome a measure of writing ability that is based on twenty independent ratings of each student's work. For the past thirty years, they have been predicting that approximately this many assessments would be necessary in order to detect significant differences between the results yielded by experimental and control treatments, but none of them, so far as I know, has ever proposed a method of securing so many independent assessments without prohibitive costs. We now have such a method firmly established in many high schools, and no one pays a penny for it above the salaries that would be paid anyway. It is not unduly burdensome, and the effect on students is bracing. This rigorous method of evaluation in itself probably produces as much improvement in writing as any of the experimental variables it is designed to study. We hoped to persuade a few high schools this year to study the effects of this measurement procedure all by itself, with no other differences in treatment, but they were all more interested is using it to study various other differences in instruction. We'll try again next year. Meanwhile, although the measure itself probably promotes growth in writing, it presumably affects both experimental and control groups to the same extent and therefore does not invalidate any other differences that may be induced by the experimental and control treatments.

After you think it over, I'm sure that many of you will have this objection. If we assume that the experimental and control groups are initially equal on the best predictive short measures we can find, the initial ratings on their test papers will presumably be equal, and they will only gradually diverge. Thus, the initial ratings really represent instruction and growth up to the beginning of the experimental year, and it will take the new procedure some months before its effects will begin to show up in the ratings. Hence the final difference between the two groups will be only about half as large as it ought to be. Possibly only the last four or five test papers should be used in comparing the performance of the two groups.

I have no objection to using only the last four, five, or six test papers as the criterion measure—as many as are necessary to attain the desired reliability. If the rating procedure to be discussed in the last section maintains the reliability it has now achieved, possibly even four such papers will be sufficient to detect a significant difference—if one exists. However, I still think it is a good idea to secure a test paper a month in order to give students practice in writing such papers and the raters practice in rating them. If the new rating procedure were introduced only during the second half of the year, it would take both students and raters a month or two to adjust to it, and the results would not be as reliable as they would have been had the monthly test paper been a normal occurrence.

Still, even if nine or ten test papers are found to be necessary to attain the desired reliability, I do not regard it as a serious objection that the initial ratings cannot be expected to diverge very much. We are content to assess the performance of a baseball team on the whole season's play, and the cases are not too dissimilar. There is little doubt in anyone's mind that the Yankees of 1963 were the best team in the American League or that the Dodgers were the best team in the National League. There is considerable doubt in the minds of us Yankee fans that the four games of the World Series proved that the Dodgers were a better *team* than the Yankees. We only grant that the Dodger pitchers were better than the Yankee hitters during this brief sample of their performance. I fear that judging the outcome of an experiment on the basis of the last few papers would be too much like judging the performance of a team on the basis of World Series play alone. I'd like to look at the whole season.

Objective Tests as Criterion Measures

Another objection that may occur to some of you is that, if the three objective tests I have mentioned are such trustworthy indicators of initial status as writers, why aren't they equally good indicators of final status? Why not give Form A of these tests at the beginning of the experiment, Form B at the end, and dispense with essays altogether?

First, because no one would be convinced. Second, because these tests have attained their high correlations with good teachers' ratings of the writing ability of their students when they have been sprung on these students once only, with little warning or preparation. If they were given at the beginning of an experiment, and if the students understood that parallel forms of the same tests were to be given at the end as the criterion of their success, it would be difficult to get anyone to write anything during the experiment. They would know that drill on reading tests, vocabulary programs, and workbook exercises on mechanics would improve their test scores far more than an equal amount of time spent on writing and revising compositions. Under these conditions of special preparation for the tests, when their specific content would be known in advance from the pretests, the validity of the final tests as indicators of writing ability would sink almost to zero. The goal of preparing for these tests would also distract both students and teachers from what they were supposed to be doing. Essays would be written and graded half-heartedly, if at all, and sincere attention would be given only to handbooks.

Why, then, are these tests good indicators of initial status? Because the specific tests chosen for any given experiment can be sprung on the students once only, without any warning or preparation—the very conditions under which they have repeatedly attained higher correlations with good teachers' ratings of the writing ability of their students than any number of essays that could be written and graded in the time available. Thereafter, they will not get in the way of the experiment, because the students will know that such tests will not be given at the end. The only way to get a high cumulative total during the experiment is to write each test paper as well as one can, and presumably to practice hard for these test papers on the three or four instructional (homework) papers that will be written before each test paper. No one will doubt the validity of this criterion.

It is not really germane to this discussion, but the question inevitably arises whenever anyone connected with the College Board is present: Will not such objective tests as these in College Board examinations also inhibit attention to writing? Our present evidence says no. These tests are also sprung on students once in twelve years of schooling with little warning as to their specific content. There are descriptive booklets, but they contain samples of all types of items that we ever use, and there are far too many for students to bone up on them. Except in coaching schools, most teachers and students are convinced that it is futile to prepare for the SAT, that the only good preparation for it is a good night's rest. It is not futile to prepare for the English composition test, since it is an achievement test; but its content differs so sharply from one form to another that it would take longer than it is worth to prepare specifically for it. Again, most teachers and students are convinced that any reasonably effective course in writing will yield scores as good as students are likely to get. Hence our questionnaires to both groups indicate that they make little preparation for this test other than becoming familiar with the types of items that are likely to be used, so that the students will know what to do. Thus, even College Board objective tests conform to the conditions under which such tests are good indicators of actual and potential writing ability. In experiments, these conditions can be met even better, since initial objective tests can be used that students have never heard of and for which no special preparation is possible.

How the Rating Scale Was Developed

In the course of this discussion, I have repeatedly referred to a rating procedure that apparently yields an initial reliability of about .7 per paper, even when the ratings are made by "ordinary" teachers who have had only one day of instruction and practice in the use of the scale. Although this finding has been replicated with different groups of raters, I still say "apparently" because the obtained figures seem too good to be true. This level of reliability is as high as we expect of our carefully selected and trained College Board readers before they begin grading a crucial examination.

We began in the fall of 1959 by getting about 600 papers from a representative sample of freshmen at Cornell, Middlebury, and the University of Pennsylvania.[1] They wrote the papers as homework between one class meeting and the next on a choice of four topics, but only two of these topics were chosen by enough students to give us a wide and parallel range of verbal ability on each topic. These two topics were "Who Should Go to College?" and "When Should Teenagers Be Treated as Adults?" There was some stimulus material on each topic. We reduced the 600 papers to 300 without reading them by dropping papers on the rejected topics and by dropping papers on the remaining topics that had been written by some of the students who had average SAT verbal scores. The result was that the writers on both topics represented a wider range in verbal ability than any one teacher would be likely to encounter in a selective college and that the distribution of verbal ability on one topic was strictly parallel to the distribution of verbal ability on the other.

---

[1] Paul B. Diederich, John W. French, and Sydell T. Carlton, **Factors in Judgments of Writing Ability.** Research Bulletin RB-61-15 (Princeton: Educational Testing Service, 1961).

We then asked sixty distinguished readers in six different fields to grade and comment on multilithed copies of these 300 papers. The fields represented were college English teachers, social science teachers, natural science teachers, writers and editors, lawyers, and business executives. These fields were chosen because we do not teach English to satisfy English teachers alone. Our colleagues also have opinions about the writing ability of our students, and so do representatives of the educated public. These readers had expressed concern about the writing ability of college graduates, and they proved their sincerity by submitting to the mortal agony of grading and commenting on 300 freshman papers for the modest honorarium of $300. We had expected that some of these busy people would be unable to complete their assignments, but 53 did—including enough representatives of each field to show that schools of thought in judging writing do not divide along occupational lines.

Their instructions were to sort the papers into nine piles in order of general merit. We did not tell them what to look for because we wanted to find out what *they* were looking for. We told them: "Use your own judgment as to what constitutes 'writing ability.' Do not assume that we want you to do this or that. Use whatever hunches, intuitions, or preferences you normally use in deciding that one piece of writing is better than another. The only rules are that all nine piles must be used, and not less than 4 percent of the papers on each topic must be placed in any pile. On as many papers as possible, write one or more comments on anything you liked or disliked."

With these free-and-easy instructions, the diversity of grades was extreme: 101 of the 300 papers received all nine grades; 94 percent received either seven, eight, or nine different grades; and no paper received less than five different grades from these 53 readers. The median correlation among the grades of all 53 readers was .31; among the college

English teachers, .41. We have been accused of stacking the cards against the essay by using such a wide diversity of readers with almost no instruction and no practice. We have been told that we should have used only college English teachers who had been trained to grade by common standards. We saw no need to do that, since our experienced College Board readers can hit an inter-reader reliability of .70 when they have agreed on criteria in advance and have practiced applying these criteria to a large number of papers before they begin grading in earnest. We were not interested in determining the reliability of an essay examination but in finding out what different things people look for when they are free to grade as they like.

The study has also been criticized on the ground that the students had no purpose in writing these papers and no clearly defined audience. On the contrary, each student was given a full account of the study, including a description of the sixty distinguished readers who would grade and comment on the papers, and two pages of printed material developing the four topics. They were stimulated by the prospect of addressing such an audience but not frightened, since they were told that their papers would be printed without identification and their grades would never be connected with their names.

I mentioned that we took pains to secure a wide and parallel distribution of verbal ability among the writers on each topic. Before discussing the findings of the study, I should mention that we found no significant differences of any sort between one topic and the other. The findings, therefore, are generalizable at least to the kinds of topics for short, expository papers that are often assigned in high schools and colleges. I do not know whether they apply to the explication and criticism of literary texts. Since writing about writing is usually the worst writing that students do, and since only one in a hundred thousand will ever engage in such writing in later life, I do not regard this exception as very important.

We correlated the grades of each reader with the grades of every other reader and with scores on three College Board tests: Scholastic Aptitude verbal and mathematical tests and the English Composition Test. We then submitted this 56 x 56 table of correlations to factor analysis. It is impossible to describe both briefly and accurately what this means. The effect is as though the computer scanned the correlations and picked out clusters of readers who agreed with one another and disagreed with other clusters to a greater extent than could come about by chance. We tabulated the comments of the three readers who stood highest and lowest on each factor to find out what they were agreeing on.

The largest cluster of readers was most influenced by the ideas expressed in a paper: their clarity, quality, and fertility; their development and support; and their relevance to the topic and the writer's purpose. The next largest cluster was most influenced by mechanics: by the number of mistakes they found in usage, punctuation, and spelling. The third largest cluster was interested in organization and analysis. The fourth stood highest in specific comments on wording and phrasing—on verbal felicity or infelicity. The fifth emphasized style, interest, and sincerity, the personal qualities of the writing, which we decided to call "flavor." The three College Board tests formed a separate factor, rather far removed from all the reader factors. I urge you not to give too much importance to this finding because different modes of assessing students nearly always come out on separate factors: grades, ratings, test scores, and the like. It is suggestive, however, that the highest correlations with the test-factor were .50 with mechanics and .45 with wording, which had a correlation of .63 with one another. It looked as though the College Board tests had attained their high correlations with good teachers' ratings of the writing ability of their students chiefly because these two factors had been the most common element in these ratings. The study, however, was not designed to investigate this point, and only ten of the readers were English teachers. Hence this finding must not be regarded as conclusive, and any inferences that we may draw must be regarded as purely speculative.

There were only three indications of occupational bias. Seven of the ten college English teachers were high on mechanics, but only one was among the three highest on this factor. Four of the seven business executives had significant loadings on organization and analysis. All of the highest readers on "flavor" were writers or editors. We could find no consistent division between men and women raters, but probably too few women were included to detect this factor—if it exists.

We next had teachers and readers in three large high schools rate papers for a year on these five factors and their principal components, and then we did another factor analysis of their ratings. It is always to be expected in such studies that the factors originally found among a wide diversity of readers in an unstructured situation will collapse into a simpler pattern under the stress of the teaching tradition and customary grading procedures. We were not surprised, therefore, to find that our five reader factors were reduced to two: a sort of "general goodness" factor and a distinct mechanics factor. These two factors are represented in our present rating card shown above. To save turning back to it, let us reproduce them here:

- General merit
    - Quality and development of ideas
    - Organization, relevance, movement
    - Style, flavor, individuality
    - Wording and phrasing
- Mechanics
    - Grammar, sentence structure
    - Punctuation, capitals
    - Spelling
    - Handwriting, neatness

It is noteworthy that four of the five original reader factors reappear as the components that had the highest "loadings" on the new "general merit" factor. The fact that they are now merged into a single factor means only that the teachers and readers were unable to make consistent distinctions among them. Students rated high on one of these components were usually rated high on the other components as well. Note, however, that "relevance," which originally belonged to the idea factor, has now merged with organization, where the teachers seemed to think it belonged. "Movement" was also added to this heading for purposes of clarification. A well-organized paper conveys a sense of *movement* toward a foreseen conclusion. A poorly organized paper, on the other hand, starts anywhere and never gets anywhere. This concept of organization resulted in more consistent ratings in our workshops than the concept of conformity to an outline.

Mechanics had the same three principal components as in our original study, but "handwriting, neatness" was added because these teachers, dealing with handwritten materials, insisted upon it. We had previously found no such factor because our fifty-three readers had dealt with multilithed copies of the original essays, but Remondino had found such a factor in Italy.[2] Our definition of points on the handwriting scale gives primary emphasis to legibility. "Grammar," of course, refers to "usage," but when we used the latter heading, in deference to the linguists, many teachers and parents asked whether we were "abolishing grammar—like those linguists." While I am not sure that linguists would appreciate the compliment, their own rules declare that we must use the term that will be most widely understood and accepted by our audience. "Sentence structure" did not appear frequently enough among the comments of our fifty-three readers to justify its inclusion, but teachers, who devote a good deal of time to it, insisted upon it. Since "grammar" includes a study of "sentence structure," we added it to that heading.

[2] C. Remondino, "A Factorial Analysis of the Evaluation of Scholastic Compositions in the Mother Tongue," British Journal of Educational Psychology, 29 (1959), 242-251.

We have little hope of getting more than two independent scores out of these eight headings, since the first and the last four generally go together, but the first four are by no means consistently associated with the last four. We might have asked for just one rating on "general merit" and another on "mechanics," but we thought that both ratings would be more reliable if we drew attention to the principal components of these factors.

After just one day of instruction and practice in the use of this scale, we have attained an average reliability per paper of .64 for the composite rating on "general merit," .63 for the composite ratings on "mechanics," and .68 for the total rating. These reliabilities are almost certain to improve as the teachers gain experience (under guidance) in using the scale. They already compare so favorably with ordinary ratings by grades that I can hardly believe them.

When we tried to introduce this rating procedure into a school that was not favorably disposed, one of the teachers objected: "All you are doing is what the Council has always recommended: give 50 percent for content, 50 percent for mechanics." That rather took my breath away. I could not remember any such recommendation of the National Council. If anything like it had ever been said by any committee of the Council, it probably took a vote rather than asking the Carnegie Corporation for $40,000, getting the cooperation of fifty-three distinguished readers, using an electronic computer, and then spending about a year in analyzing and tabulating comments. I fervently denied that the four principal components of the "general merit" factor in its present form were adequately described by the term "content." It is not a triumph of common sense over science that these four components have temporarily blended into a single factor. I look upon it rather as a defeat: The teachers, bound by the teaching tradition and conventional grading practices, were unable to make the number of distinctions that we asked them to make. As they become more adept in the use of the scale, I expect at least a few of these ratings to emerge again as independent factors.

After we had answered a number of such objections, the final objection of the English faculty of this school was that we were interfering with their constitutional right to grade however they pleased. Were they all supposed to think alike? What would become of their inalienable right to free thought and free speech if they all had to grade on factors that came out of a machine?

It would be fruitless to try to convince a faculty like this; we merely said goodbye and went on to another school. But if there are any lingering doubts on this issue among members of this conference, let them remember this:

The machine made no decisions whatsoever; it only discovered agreements and disagreements among our fifty-three distinguished readers. As a research team, we made no decisions other than trying to figure out what our readers meant by what they said. When we went into this study, we had no idea what might come out of it; and we took precautions that would be inconceivable to the unscientific mind to keep our private opinions from influencing the outcome. What we have to offer now is simply a picture of the elements in student writing that are valued by many distinguished people—in our own field and out of it—who care about writing. Then we tried for a year to get English teachers in three large schools to rate student papers on these elements, and we reduced our original list to eight elements that they were able to distinguish reliably. If this is tyranny, what is freedom? At no point have we tried to interfere with the thought processes of readers; we have only tried to study what in fact those processes are. If we now tell you how distinguished people think about student writing, and what they value in it, you are under no compulsion to think as they do. All you have to worry about is that some day your students are going to write not for you but for other people, and then the writing of your students is far more likely to be judged by the standards we have discovered than by any private and personal standards that happen to occur to you at any given time. If you accept our standards during an experiment, you can find out something about the teaching of writing; but if each man is to be his own judge, no conclusions are possible on which any two people will agree.

RESEARCH DESIGN IN ELEMENTARY SCHOOL READING AND LANGUAGE

Donald D. Durrell

Since research design depends upon the nature and complexity of the problem studied, it seems well to begin with a look at some problems to be undertaken in reading and language. Research design, as I conceive it, is the management of the variables under consideration, not the relatively minor problem of statistical treatment of data.

Before undertaking this paper, I engaged in the chastening process of summarizing the 500 masters' and doctors' theses in elementary school reading done at our institution during the past thirty years.[1] The yield of answers is remarkably small; the yield of new questions is the main harvest. Our fire was too broadly scattered for effective results. In the field of reading readiness, we dealt with problems of age and mental age for school entrance; with the evaluation of reading readiness tests and workbook activities; with phoneme perception and letter-name knowledge; with experience backgrounds of children, including kindergarten attendance; with the construction and evaluation of tests and teaching materials. In primary grades there were many studies of phonics and word recognition; of speaking vocabulary backgrounds; of abilities in oral reading, of comprehension and recall; of pupil interests and activities; of textbook and test analysis; of evaluation of teaching materials and techniques. In intermediate grades, emphasis was upon the following: word skills, including analysis, meaning, transfer skills, and relationships to spelling; problems of sentence comprehension; the analysis and improvement of recall; a variety of study abilities and higher mental processes; imagery in silent reading; reading interests and habits; textbook and test analysis; evaluation of materials and methods. Looking back at the product, I am sure that we would have been more effective if we had concentrated our studies in fewer areas.

Scattered Research Designs

Research designs, of course, varied with the problem and with our knowledge or ignorance of the variables impinging on the problem. Some

---

[1] Donald D. Durrell and Helen A. Murphy, "Boston University Research in Elementary School Reading: 1933-1963." Journal of Education, December, 1963.

designs were relatively simple, such as a study of letter knowledge brought to first grade. This involved building tests of different levels of letter knowledge, selecting a suitable population, administering the tests, and analyzing the results for various segments of the population. The study of speaking vocabularies of children had the usual problems of studies of its type: provinciality of sampling, variety of stimulus situations, methods of recording and analyzing protocols. Studies of imagery in silent reading required designs to meet the hazards of introspective techniques, especially when used with children. Test construction and evaluation used the familiar designs of item analyses and estimates of validity and reliability. Status studies, such as surveys of error types, sex differences, age-grade differences, and similar analyses required relatively simple procedures. Studies of the influence of various perceptual factors on reading utilized techniques of partial correlation or pairing out of variables; possibly we should have made use of factor analysis.

When one enters the field of evaluation of methods and materials, the multiplicity of variables forces more complex research designs. When our methods studies involved discrete learnings, such as the spelling of word lists or recall of particular groups of facts, rotation methods were used to control pupil and teacher variables. Each learning unit was taught by different techniques, with planned rotation to involve all pupils and teachers in each technique. Results from different techniques were then tabulated and compared. Simple studies required only 2x2 rotation, while the most complex one used both 8x8 and 12x12 rotation plans. When we studied cumulative effects of different procedures, as one must when dealing with transferable abilities rather than discrete learnings, we used before and after designs, controlling pupil and teacher variables as best we could by matching groups or individuals on variables which seemed to be most pertinent. The Hawthorne effect was minimized in some studies through comparison of two or more experimental approaches which seemed equally promising. However, too many of our studies compared a single experimental approach against "the standard approach," which often lacked definition as well as Hawthorne potential. Incidentally, our experience in methods research does not bear out the canard, "In methods research the experimental method always wins"; many of our bright ideas didn't win.

This summary of scattered research designs is intended only to illustrate the obvious point that the research problem determines the design. I think that one of our major concerns here should be a consideration of ways to increase the yield of knowledge from language arts research. For many years all of us have directed independent studies in minor problems of reading and language arts, yet I think all of us are disappointed with the general growth of dependable knowledge from these many studies. Part of the difficulty may be in inadequate dissemination of research findings. Many people have proposed language arts research retrieval centers, replete with electronic devices to answer questions. Somehow this proposal leaves me with doubts; the questions will be complex and the answers clouded with verbal haze. Another difficulty may be flaws in research design. However, I doubt that improvements in the techniques of isolated research skirmishes will move us forward greatly. What is needed are new strategies for a long and complex research campaign.

Components of Language Arts Learnings

One of our first tasks is that of development of bodies of knowledge concerning the structures and functions of language arts learnings, something akin to the fields of anatomy and physiology in medicine. We lack precise knowledge of the nature of the components to be learned in language arts. While we have a crude gross anatomy of reading, listening, and writing, we are vague about most of the subskills to be learned in each area.

The need for subskills analysis is well illustrated by the problems encountered in beginning reading. The task appears to be that of matching phonemes to their printed forms. Language studies provide us with a knowledge of English phonemes, and we know the letters of the alphabet. When we begin to teach reading, however, we meet all sorts of difficulties because of lack of specific knowledge of the subskills encountered. What level of awareness of phonemes is needed for beginning reading? How do children differ in their knowledge of the sound patterns of spoken words? Is it a general power, or must one take an inventory of identification of individual phonemes? Does the ability to identify the phoneme at the beginning of a word assure its recognition in medial or final positions? Is there an order of difficulty in phoneme recognition? Will the order be different when the juxtaposition with other phonemes varies? If one may judge by professional textbooks in linguistics and reading as well as by many phonics programs, many of our colleagues are still unaware of the need for specific ear training as a basis for beginning reading.

Letter Recognition

Moving from the awareness of phonemic patterns in spoken words to their attachment to letters and words, many more problems are encountered. What degree of letter recognition is necessary for ease of learning? Is the ability to match letters enough? Our studies show that the knowledge of letter names in September predicts June first grade reading achievements better than most measures. But why? Does it indicate a higher familiarity with letter forms, or is its merit in attaching sounds to the letter since most letter names contain their sounds? Is the ability to give letter names an early stage of phonics? Or does letter knowledge have no value as a perceptual element in learning to read, its predictive value resting only on its indication of general perceptual maturity? Should letter names be taught as the earliest childhood reading experience, or should the attachment of sound to the letter be first? Should not the writing of letters be the first approach to reading, since the desire to write appears before the desire to read? Does the difference in forms of letters in print, manuscript, and cursive writing make much difference to the beginning reader? Although many of us have our own answers to questions such as these, the evidence we have presented seems not universally accepted.

Spelling

To complicate the problem further, how much trouble results from the fact that phonemes are spelled differently in English? Shall one teach the variations immediately, showing common spellings for sounds, or shall we teach the "dependable" sounds exclusively at first, using the *Fat-Sam-sat-on-a-mat* approach advocated by some linguistics anatomists? Is it necessary to provide a pure phonics approach with new symbols or variation of current symbols to indicate each phoneme? How about various colors of print to indicate phonic changes? And shall we teach phonics apart from meanings, or shall we intensify meanings in phonics teaching? Do we teach phonics rules, or shall we let the child acquire his generalizations inductively? Since many children are unable to synthesize the separate sounds they produce in word analysis, should we not teach phonics by an analytical look-say process? How much value is there to sounding out of words? Is there any point to teaching short and long vowels, or are these better learned in phonograms and syllabic units?

What will be the effect of each of these approaches on spelling, since spelling introduces new functions different in many ways from reading? What abilities do good spellers have that are not possessed by poor spellers? One of our studies of homophone recognition showed that good readers and spellers can identify various spellings of the same sound; the test correlated .80 with spelling. Does this mean that poor spellers should be taught homophones directly, or will this only result in a greater variety of misspellings? Will a high meaning-imagery approach be desirable for teaching all words as it appears to be for homonyms? Does the correct spelling of a word in one meaning assure correct spelling of the word in other meanings? How do we assure transfer from spelling to writing? Shall we cluster words by structure or by meaning? At present we cannot claim to have adequate knowledge of the subskills of spelling or sure techniques for improving achievement in it.

Meanings in Context

We have thus far considered only single word functions, ignoring the sentence and the problems of meanings in context. The pronunciation of homographs often depends upon context, but this is a minor problem. What effect does sentence organization have upon comprehension? Must sentence patterns in reading follow those in children's speech? Must we test each type of sentence structure for difficulty and provide an orderly sequence for teaching? Or can we assume that extensive reading will care for this inductively? Children differ in their ability to acquire sight vocabulary and new meanings from extensive reading. What is the nature of the ability to grasp and hold new words from contextual reading? Must one teach directly the ability to notice new words and to utilize a variety of context clues? And since the signals in speech provide richer meanings than the same sentence in print, shall we develop new ways to transfer this factor to silent reading?

Before we move to more complex language arts abilities, we should perhaps consider whether or not these questions need to be answered. Do we need such detailed anatomizing of subskills? Do not millions of children learn to read despite our lack of such knowledge? Of course they do, just as there are millions of people who would maintain good health if there were no knowledge of anatomy and physiology. But certainly the general health of populations improves with application of principles of health; and in illness, such knowledge is essential. We may have the same expectations for improving knowledge in the anatomy and physiology of language learnings. Until we have better information we shall continue to be vulnerable to the panaceas of cultists and the practitioners of educational and psychological witchcraft.

Precision in Knowing Subskills

Perhaps it is unnecessary to provide more illustrations of our need for more precise knowledge of the subskills of language arts. We know that reading is much more than the recognition of words and the reading of sentences. We know too little of the nature of intake of ideas from reading. The essence of reading fiction would seem to be the imagery it evokes; but the analysis of the types, amounts, and quality of images from reading is relatively untouched by research. What are the factors which aid or impede the understanding of materials read? What are the conditions which improve speed of silent reading? Can every child be taught to read five thousand words a minute without loss of comprehension? In the areas of recall, we know that great differences exist among children of the same mental age and reading ability, but we have not begun to study the reason for these differences. In fact, we have no satisfactory measures of oral and written recall; we appear content with the objective test level of identification of correct answers. When it comes to the use of knowledge from reading in various thinking abilities, we have only vague beginnings on the problem. We speak of critical thinking as though it were a unitary ability; we have little knowledge of various types of elaborative thinking; we have few useful measures of gross abilities in thinking and no truly analytical measures. In the areas of study abilities, we have the long lists of presumably discrete abilities found in readers and workbooks, yet we do not know which are essential or which are truly discrete. This, of course, brings up the problems of transfer. Surely we will find that some study abilities have common elements that will readily transfer, and we can focus our teaching on these. It is also necessary to study the problems of interest and initiative in voluntary uses of reading. There is little advantage in teaching an ability that dies from disuse.

We all recognize that the need for similar detailed analysis of structure and function is even greater in other language arts. Only in the field of clinical speech have we the beginnings of adequate professional knowledge. The subskills of written composition are still largely unknown, except in the relatively minor area of language-grammar correctness. We are still seeking panaceas through such approaches as the use of lay readers, motivation techniques, smaller class size, teacher-team organization, television presentation, personality attributes of pupils and teachers, *and* that perennial hope—curriculum revision. We need intensive studies of the processes involved in composition and their responsiveness to various teaching approaches: such as the flow of ideas related to a topic; sentence variety; organization, sequence, and transition; richness and precision of vocabulary; variations of style and emotional tone. Composition probably will be found to have a greater variety of subskills than any other language arts area. But we should try intensive teaching of some of the known subskills, recognizing that we shall still have the problem of transfer to the whole writing task. Perhaps a subskills approach will provide the child with knowledge of progress, making an unwelcome task more acceptable. Studies of subject preferences among intermediate grade children always find language arts the lowest of subjects, and writing tasks the lowest item in the language arts.

The power of subskills analysis and instruction is best seen in individual tutoring. In tutoring, we can discover subskills weaknesses and direct our attention to them; we can also adjust to level of ability and to learning rates. Intensity of learning is also high, since all instruction is focused upon the needs of the individual. When we move to class instruction, much of this precision of adjustment is lost. We need to search for ways to maintain the specific subskills services in group instruction, especially through inclass grouping and the use of self-directing and self-correcting materials.

If we are to develop a body of secure knowledge in the subskills of language arts, we must adopt new research plans. More of us must concentrate on narrower areas and stay with them year after year, as do our colleagues in biological and physical sciences. We have constant pressure for broader scholarship in the education of our students; perhaps it is time to think of narrower concentrations of scholarship. No person can possibly be a general scholar in the field of language arts.

Research Strategies

Three research strategies are suggested for increasing the yield from language arts research: long-range careers in sequential studies in limited areas; large-scale intensive studies in limited areas; coordinated studies at a number of centers which deal with a relatively narrow instructional field. The first can be followed by the individual scholar who limits his field and pursues it for life. The second requires large grants for multiple studies in limited areas, with sufficient time for redesign of research after first attempts.

For the first time in educational research history, we have the opportunity to try the third approach: that of coordinated studies using many different approaches in a limited field. This is provided by the invitation of the Cooperative Research Branch of the U.S. Office of Education to submit proposals in beginning reading instruction. This is the culmination of research conferences of the National Conference on Research in English at Syracuse University in 1959 and at the University of Chicago in 1960. The plan allows each center to follow any approach it wishes; each should evaluate two or more approaches; each may use unique measures for solution of its particular problems. However, there will be coordination of the studies through conferences of principal investigators of accepted proposals. They will decide upon methods of control of variables, the selection of common measures, and techniques of analysis of data. We do not yet know how many centers will be selected, nor the precise bases for selection of centers. Some consideration should be given to the balancing of projects with highly contrasting approaches which will yield the greatest amount of basic information. It is regrettable that the date of submission of proposals is late in the year; it will be difficult to handle the complex task of coordinating the studies in the limited time available after the projects are selected. A research development seminar devoted to this single study should be initiated as soon as possible.

There are many advantages to coordinated studies in a single area. Many minds are brought together for research design; each separate study will have the advantage of the research experience of other scholars. Instead of cozy noncomparable studies yielding conflicting findings, the coordinated study promises comparable studies and findings. Surely there will be a greater stimulus to intellectual effort when the researcher is aware that others are working at the same problem. And certainly the impact of the coordinated study on educational practice will be far greater than that of a separate study. It is also easy to find difficulties in coordinated studies. A single methods study is difficult to control and evaluate; simultaneous methods studies will be even more complex.

The yield in basic knowledge and improved school practice might be very great if we conducted coordinated studies of written composition in grade nine; of spelling in grade five; of critical and elaborative thinking in grade seven; of programs for superior learners in grade six; of English programs for students in grade ten. Admittedly one needs longitudinal studies to discover cumulative effects, but it would be interesting to see the yield from such limited studies.

RESEARCH DESIGNS NEEDED IN STUDYING THE TEACHING OF LANGUAGE AND LITERATURE IN THE SECONDARY SCHOOL

Henry C. Meckel

Let me focus attention first on two matters that I consider to be of major importance in the planning of research design: the need for a rationale of learning that is specific to our field, and the need for investigations that focus on sequences in learning. After discussing these two topics, I shall suggest certain areas for research that appear to be significant for the teaching of secondary school English.

The Need for a Rationale of Learning in English

Much attention is being given today to the problem of controls in research design. The difficulty of controlling experiments has not only been a plague upon those attempting educational research but has also been a factor discouraging psychologists from studying human learning. A good friend of mine suggests that awareness of the importance of experimental design has been responsible more than we realize for leading psychologists to investigate learning with rats, chickens, apes, or computers as subjects. Such experiments are much easier to control, and they fit the kinds of rationale that have been developed as learning theory. Moreover, in the hands of a clever and imaginative investigator they can produce fascinating results. The results, however, are not always helpful to people concerned with the process of educating human beings in classrooms.

We should recognize, therefore, as we wrestle with the problem of experimental design, that in attempting to study teaching and learning in the classroom we have elected to step where psychologists have often feared to tread. Remembering this should help everyone concerned to be patient. Studying human learning takes time.

My acquaintance with studies in the teaching of language and literature suggests to me very strongly that we cannot adequately study the process of education in English without a rationale and a theory of learning as a basic frame of reference. The report from the Research

Development Seminar at New York University suggests that my feelings may be shared by many other people.

Such a rationale of learning must be specific to our area of study. One of the most significant results attending the enthusiastic reception accorded Bruner's *The Process of Education* is the emphasis given to the importance of the structure of knowledge in learning.[1] Learning theories have in the past been of a generalized character; they have not incorporated the idea that the content learned has within itself characteristics that affect the learning process. Learning theory has treated the learning of concepts, skills, interests, attitudes, and appreciations without reference to the nature of the subject matter to be learned, just as it has dealt with the problem of transfer in a general way.

The relationship of learning theory to research design is evident in the paper by Campbell and Stanley on experimental design published as Chapter 5 in the *Handbook of Research on Teaching*.[2] The authors of this paper strongly feel that the spurious nature of much educational research lies in the unsatisfactory psychological orientation of experimenters. They suggest that research designs become cluttered and unmanageable when they are not grounded in some form of behaviorism.

It is my belief that much educational research, especially research in the teaching of language and literature, has failed to be of significance largely because the hypotheses have been formulated in simple stimulus-response terms. The completed studies have been limited in significance because from the very beginning they were based on fallacious conceptions of human learning and educational method. Some examples from past investigations in the field of English will serve to illustrate my meaning. I will indicate the nature of the studies by stating the questions which guided the investigations.

Do pupils who have been taught grammar write better than pupils who have not? If pupils have more practice in writing, do they write with more skill? If pupils choose their own composition topics, do they write better than pupils who write on topics assigned by teachers? Do pupils who read literature extensively learn to interpret more or less adequately than pupils who study literature intensively?

The hypotheses guiding these studies seems to be based on a simple relationship between stimulus and response. With such an orientation it is easy to overlook essential phases of educational method and learning process and to concentrate attention on peripheral matters.

A study to test the usefulness of grammatical knowledge with respect to composition should be based on a careful consideration of the grammatical concepts that can be expected to be most functional. This is a problem that presupposes the definition of a structure of knowledge. Such a study must also consider kinds of learning experiences that will be conducive to transfer. Sequences in learning and teaching need to be carefully considered. A rationale for learning of grammar needs to be formulated which will incorporate hypotheses concerning optimum prerequisites for transfer. The rationale will also have to include some reasonable suppositions with reference to learning time. Most research studies in grammar have generally involved too few training sessions to develop understanding, too little attention to the conditions necessary for transfer, and almost no careful consideration of the learning sequences necessary.

---

[1] Jerome Bruner, **The Process of Education** (Cambridge: Harvard University Press, 1960).

[2] Donald T. Campbell and Julian C. Stanley, "Experimental and Quasi-experimental Designs for Research on Teaching," in N. L. Gage (ed.), **Handbook of Research on Teaching** (Chicago: Rand McNally & Company, 1963), Chapter 5.

My second example concerns practice and skill in composition. An investigation having to do with the relation of practice to the development of writing skill cannot be educationally significant if the design is developed in purely behavioristic terms and quantity of practice becomes the variable to be investigated. A significant investigation must involve hypotheses concerning learning sequences based on some rationale delineating the phases of teaching method most likely to provide favorable responses in order that there be favorable opportunities for practice to contribute to skill.

My third example, studies based on the hypothesis that pupils write better when they choose their own topics, illustrates research design that involves a rationale of learning, but not a rationale that includes a structure of knowledge. Such studies usually are based on a theory of learning which gives prominence to the assumption that pupils are more motivated if "authoritarian" elements are at a minimum in learning situations and if pupils have maximum opportunities for choices. Another assumption present in such a theory is that pupils are more interested in what they decide to do than in what the teacher plans. The research design treats teacher-selected topics or student-selected topics as stimuli that will produce uniform responses. At such a simple behavioristic level, the theory of learning actually permits the formulation of hypotheses that are not educationally significant. Effective assignments are of course an important phase of good teaching. They are most fruitfully investigated, however, when they are regarded as a phase of learning sequence in which readiness for learning is an important concept. In this connection I am thinking of readiness, not as a factor associated with maturity, but as a prerequisite for a learning task—a prerequisite which the teacher may facilitate. The nature of the sequence involved is not likely to be adequately described and the aspects of readiness determined unless the elements of knowledge—the concepts and skills involved—are clearly conceived.

The Need for Investigations in Sequence of Learning

Let us consider research with respect to the extensive and intensive study of literature. A good program of instruction in literature probably should involve both intensive reading, each having an essential place in a carefully planned learning sequence. The treatment of extensive or intensive reading as methods, rather than as phases of sequence in method, can lead to studies which cannot be convincing to good teachers, however adequately we may specify controls in setting up design.

I am emphasizing by these examples the idea that appropriate sequence is an essential characteristic which distinguishes good method from inadequate method. I suggest also that it is ridiculous to consider a lecture method, a discussion method, a laboratory method, or a television method as if these "methods" existed apart from the subject matter or learning sequence in which they are used. If we plan experiments in which these procedures themselves are hypothesized as producing measurable effects, we will not achieve results of true significance. Between the stimulus and the response are human learners with predispositions and knowledge and skills, and they do not function in the same predictable way as animals or computers. This seems to me to be a fact that we must face, and it makes the planning of research in our area extremely difficult.

In recent papers on research in English, I have not noticed the problem of sequence recognized in the sense I have been suggesting, except in a proposal for an exploratory study of prewriting techniques suggested by Carroll and Rosenblatt at the New York Conference.[3] Other proposals at the same conference seem to imply the idea, however. In most projects, interest in sequence seems to be limited to curricular sequences of instructional materials at different grade levels, as though there were some absolute answer to the grade placements of particular works of literature which research or the opinions of professors can once and for all determine. But whether a specific work of literature or any other content is suitable for any grade level or a particular group of pupils at a grade level often is a matter of what the teacher does with it, and what he does is frequently the result of the sequence of instruction that he plans.

I should like next to hypothesize concerning the nature of sequence as an essential characteristic of method. This is an area of learning about which we know very little. We can, however, hazard descriptions about sequence based on what teachers do.

---

[3] See Proposal No. 3 in Louise Rosenblatt (ed.), **Research Development in the Teaching of English**. Project No. 6-009, Cooperative Research Branch, U. S. Office of Education (Washington: U. S. Department of Health, Education, and Welfare, 1963), p. 47.

Let us consider first a relatively simple example of language skill, skill in writing an expository paragraph. Let us suppose that sequence will involve the following elements: First, the pupil must understand what a paragraph is. (This understanding is a concept of a language structure.) Then the student must learn principles of paragraph development—the principles that apply to unity, including the idea of pertinent and extraneous details; the methods of expanding, explaining, or supporting an idea; the methods by which sentences within the paragraph may be related to one another. (These are principles of structure.) It is not sufficient that pupils merely be told these principles or that they be able to state them. The principles have to be established in the mind of the pupil with the special kind of reality that comes with experience and use. This kind of understanding not only entails writing but also using the principles to analyze the content and structure of paragraphs which the students read and study. In this phase of the sequence, establishing relationships between reading and writing is important. The pupils must also have sufficient practice in writing, revising, and polishing their own paragraphs to gain a degree of skill that approximates habitual, independent performance. In this hypothetical description the words *concept, principles, application, experience, interrelationships of reading and writing, habit,* and *independence* suggest some of the elements involved.

My next example has to do with the factor of *readiness*. By readiness in this sense I refer to the types of skill that must be developed by a teacher in order that students may succesfully complete a major assignment. Let me illustrate. A teacher of a superior class decided that as a major assignment each student would read one of Plato's dialogues very carefully and write a critical analysis of the argument. The sequence of learning activities was as follows: (1) The teacher assigned a Platonic dialogue and led the class in a critical analysis, thus demonstrating the approach and techniques required. (2) Knowing that his students were still not ready to do the main assignment, the teacher assigned a second dialogue and arranged to have the argument analyzed in small groups. At this stage the teacher thus arranged for independent practice, contributing to the security of students by making himself available as a consultant and by encouraging diverse students to contribute to the learning of the group. (3) The teacher assigned a third dialogue, which students analyzed by themselves, each writing a paper independently. This illustration suggests that there are such additional elements in a sequence of learning as the *development and sharpening of conceptualization, security in initial stages of practice when new principles and concepts are involved, increasing awareness of the relation of principles and concepts to performance, and a progressively higher level of independence in learning.*

It is extremely difficult to conduct experimental studies of a sequential process of human learning. If the researcher attempts to do so over the period of time represented by a curriculum, a course, or even a unit of instruction, many uncontrollable elements are likely to be present in the research design which restrict the extent to which the investigator can generalize about sequence itself as an element in instruction. If the focus is narrowed to one or even several class periods, the investigator is likely to find himself with an isolated moment or a segment of a process which is of too short a span to be significant or which involves outcomes too minute to measure.

It is better, however, to make beginnings and attempt to secure some data on which to base opinions than to indulge completely in guesses. We can attempt to learn how to describe the process even though we cannot generalize about the precise relationships of the various elements within it. I would, therefore, like to see integrated courses in literature and composition developed, in which sequences of instruction are carefully planned, not only in terms of literary works to be studied and concepts and principles to be learned, but with attention to the various abilities of particular groups of pupils and their levels of readiness. If no other type of research is practical, I would be satisfied for the time being with descriptions of results in terms of gains made on standard tests, with appraisals of samples of writing included as additional supporting evidence. I would also want to include in the evaluation the teachers' opinions of the course. In addition, if teachers were asked to indicate six students who benefited the most from a course and six who benefited the least, interviews with these students, together with cumulative file data, might lead to better understanding of the results of instruction. This would be mere descriptive research, but it is not without merit.

Further Possibilities for Research

A design for a study of a specific phase of sequence has been suggested by Carroll and Rosenblatt.[4] The phase of sequence involved is "prewriting" instruction and pertains to those techniques which a teacher may use to stimulate students' thinking in preparation for writing. A strong feature is its simplicity and adaptability to classroom situations and its possibilities for training teachers to plan and evaluate phases of instruction that have sequence. Moreover, the proposed design is capable of being replicated in enough situations so that a research center associated with a number of field schools could collect and treat data from many teachers and produce a respectable study.

---

[4] Louise Rosenblatt, op. cit.

The design suggests the possibilities of forming four groups within a class by random assignment of pupils—four experimental groups and four control groups—and using samples of writing as pretests and posttests.

An exact duplication of this design could be used to test the effectiveness of various theme revision techniques, such as checklists and small group discussions, which are commonly used by teachers to assist pupils in revising first drafts as a step in producing good papers.

The design could be modified for other manageable classroom investigations of sequence if, through the efforts of a research center, classes in cooperating schools were organized on the basis of random selection and teachers assigned on the same basis. Teachers' manuals outlining very specifically the steps to be followed in an instructional sequence would be provided, and teachers participating in the experiment would be carefully briefed in order to control conditions of instruction as much as possible.

This general design might be used to test the following hypotheses:

1. That it is possible to arrange the reading of short stories in one sequential order that will be more effective than another in developing ability to evaluate literary craftsmanship.

2. That in studying the short story, listening to recordings of stories by professional readers may be more effective in developing an understanding of style than reading stories silently and talking about them.

3. That instruction aimed at developing understanding of such techniques as characterization, point of view, or plot structure is more effective if it includes the following sequence: (a) teacher explanation, (b) student analysis in small groups, (c) individual analysis in a paper. (What I am proposing here is a small study of progressive independence in learning. Additional experiments might be directed toward discovering how many times such a sequence must be repeated; whether, if a programed sequence on concepts and principles were inserted between steps (a) and (b), the total sequence might be improved; or whether high quality teacher explanation, perhaps introduced in well-prepared kinescopes, might improve the effectiveness of the sequence.)

4. That a sequence of instruction involving a literary concept such as

symbolism, mood, tone, or style, for example, is more effective if at initial stages an attempt is made to clarify the meaning of the concept through very concrete examples and the use of audiovisual aids.

5. That a short story unit which involves extensive supplementary reading of stories (reading which is recorded but which does not involve book reports, say) is more effective than one which does not require reading beyond the anthology.

Sequence experiments, from the point of view of a particular school, may be quite modest, but, when replicated in a number of schools under the coordination of a research center, they may be respectably extensive. Such projects may have special advantages in training teachers in the idea of sequential teaching. They would give teachers experience in the progressive examination of the results of their teaching when, within a sequence, different outcomes are demanded at different stages; for example, understanding of a concept, application of principles in an assignment requiring analytic reading to examine structure, or putting newly learned principles to work in the initial stages of teaching composition. Such projects, therefore, may be the means of developing sophistication about learning. Detailed manuals for short instructional sequences might be prepared with relative ease and would have the effect of standardizing procedure for experimental purposes. Pretesting and post-testing present a special problem, however. In the absence of objective tests, perhaps essay tests or other pupil writing might be used, and scored by means of rating procedures according to categories that would describe the degree of understanding of a major concept or the quality of application of principles which pupils make to a reading or writing assignment.

In another area of research in the teaching of English, designs are greatly needed to study the process of interpretation of literature. English work has been done in this field to indicate that patterns of response in reading are frequently of such an individual character that a stimulus-response model is not adequate. Instead of a simple relationship, there is, between stimulus and response, a constellation of predispositions that are symbolic, emotional, and personal in nature.

Patterns of response are probably specific to different literary genres. They are likely to be most personal in fiction or drama in which there are parallels between the experiences of the reader and the experiences of characters. More information about the process of identification with characters and situations, for example, would help teachers to understand the dynamics with which they must deal in promoting what Rosenblatt has called "transactions" with books. Good studies require subjects about which a great deal is known and insights from men trained in psychoanalytic psychology so that motivations and predispositions to response are validly interpreted and investigated. Very careful decisions must be made with respect to samples of materials used for reading and interpretation. If statistical procedures are used, advice is needed on techniques most appropriate for treating data for small groups. If more were known about pupils' responses to fiction and poetry, more precise statements might be made about the sequences of instruction most likely to be effective in modifying responses. In addition to developing designs for investigations, we need to develop categories not only for the analysis of responses but for the analysis of literature that is likely to evoke them.

Possibilities for developing research designs for studying the teaching of English language and literature are continually affected by the state of measuring instruments available. Research directed to the development and improvement of tests is therefore especially important. Strong endorsements should be given to proposals for the development of tests on specific literary works and for the production of equivalent forms of such tests. Tests for measurements of components of composition skill—in addition to those that attempt to appraise skill broadly—are needed if phases of instructional sequence are to be investigated.

In closing, I should like to point out that we do not have data in the area of writing skill available which suggest the level of writing ability that may be expected to be associated with average intelligence, as measured by nonverbal tests, or with average intelligence as associated with different socioeconomic levels. We do have data of this nature in the field of reading. I would like to suggest that there is need for such data to serve as a reference point for the levels of performance we expect of high school students. Levels of skills required for college and university work may be very high if we take them also as the levels of proficiency which we can attain with other types of students. Without in any way receding from our responsibilities to develop programs which prepare college bound students for their educational tasks at university and college levels, we may well need to give our attention to some research designs associated with investigations in which attention is focused on the skills needed by the terminal student.

It is easy to detect the flaws in research already done, but it is difficult to design projects which will not do violence to the process of English instruction. So often a research problem that appears on the surface to be the simplest kind turns out, when one begins to work on details, to be full of hazardous procedures. Every suggestion I have made in this paper is of this nature. In research in English instruction, except perhaps in composition where large amounts of money have been spent recently on improving measurement, we are not in a position at present to do research which will result in generalizations with wide application. We know next to nothing about the nature of pupil responses in the phases of learning that are really significant in English, we disagree among ourselves about the nature and structure of our subject, and we have only the clumsiest instruments of measurement at hand. Until we make more progress, research in the teaching of English at the level of the secondary school is going to have to be largely exploratory and descriptive. That such is the case should not lead us to be careless about designs or to avoid facing up to problems. Psychologists and psychometrists cannot be of any help to us unless we ask questions that are really pertinent to our field.

III. SOME TECHNICAL APPROACHES

The three papers in this section discuss and illustrate some techniques with which to attack research problems in the teaching of English. As both Gage and Diederich indicated in earlier sections, the simple design of comparing two methods of teaching some phases of English by the comparison of gains in an experimental and a control group has seldom yielded differences which allow generalizations about teaching methods. Some other patterns for research are needed. The papers in this section are examples of other types of analysis or improved procedures in the parallel groups pattern.

Mitzel points to the importance of exact description of classroom behaviors if an experimental and a control group are being compared. English teachers or research workers will need to apply his general ideas about recording activities to specific types of language behavior.

The papers by Harris and Peel discuss variants and uses of factor analysis. The Harris paper is of special interest to design specialists who go beyond simple correlational studies to discover patterns in relationships. The paper offers one example of a place where the English specialist must depend upon the competence of the design specialist in deriving results which both persons, working together, can interpret. The Peel paper is an interesting example of going beyond rather superficial studies of child or adolescent reading interests (many of which are content to list only favorite books or types of stories) to determine underlying factors related to literary preferences or appreciation. He illustrates a way of measuring appreciation of writing in terms of certain qualities in the material as rated by competent judges. It is difficult to get at causes of liking and disliking, but Peel feels justified in going so far as as to use the term *determiners* of reactions to visual art, to poems, and to children's compositions.

THE PLACE OF SYSTEMATIC OBSERVATION IN CLASSROOM RESEARCH

Harold E. Mitzel[1]

This paper presents a discussion of the role of systematic observation as it functions in general classroom research. I hope that what I say is applicable to research on the teaching of English.

Systematic Observation Defined

I shall confine my use of the term ‘observation” to its meanings in psychological and psychophysical research. There we use it to describe perception employing any one or all of the five senses, and, in addition to its meaning as a process of perceiving, we use "observation" to describe a single piece of the data obtained for scientific study. To my knowledge, psychologists seldom use the term in its original meaning, i.e., "to observe a ritual."

By use of the term "systematic," I intend to disqualify those unplanned, haphazard types of classroom observation that are usually made for the purpose of supervision by department heads, school principals, and others who bear a line or staff relationship to the classroom teacher. I also classify as unsystematic, observations that are predicated upon the model of the sponge. Thus, an observer who maintains that his goal is to visit a classroom to "soak up" the atmosphere or the feeling-tone, or the quality of the instruction, is engaging in an unsystematic activity. To be systematic, the observer should know in advance what behaviors, events, or activities he wishes to observe. The second hallmark of systematic observation besides its planned quality is its record yielding characteristic. The systematic observer in the classroom makes a record of the behaviors, events, or activities he wants abstracted from the total milieu. This applies whether he observes directly, or elects to make sound tape, videotape, or kinescopic recordings of classrooms in action and takes the relevant data from these secondary sources. In any event, some sort of scientifically manipulatory record is obtained for further study.

---

[1] In addition to those given in the footnotes in the following presentation, the author cites these references:
A. A. Bellack, **Theory and Research in Teaching** (New York: Bureau of Publications, Teachers College, Columbia University, 1963).
D. G. Ryans, "Assessment of Teacher Behavior and Instructions," **Review of Educational Research**, 33 (October, 1963), 415-441.
H. F. Silberman (ed.), "A Symposium on Current Research on Classroom Behavior of Teachers and Its Implication for Teacher Education," **Journal of Teacher Education**, 14 (September, 1963).

In talking about classroom research, let us agree to include studies of school age youth in group situations as well as research in formally organized classes. It might be well to recognize that the principles and generalizations stemming from classroom behavior research have the potential for early obsolescence. Team teaching and instructional formats utilizing programed text materials, computer guided instruction, and television are all capable of challenging the ubiquitous classroom composed of one teacher and some thirty-odd pupils. The history of education from Socrates to Skinner can be viewed in terms of a gradual shift away from the teacher as the single source of information for the pupil. Inexpensive textbooks provided the teacher with the first important auxiliary for information transmission, and the subsequent widespread use of texts in schools dictated a change in teacher behavior patterns. Newer educational media, insofar as they receive enthusiastic adoption and provide more channels of information for the pupil, are going to modify further the behavior patterns of teachers. In a sense, we are much too close to the revolution in education to be able to perceive its course or even its eventual impact on the process of schooling.

Characteristics of Careful Observation

In the American Educational Research Association's recently published *Handbook of Research on Teaching,* Medley and I have reviewed studies from the educational and psychological literature which employed techniques fitting our definition of systematic observation of classroom behavior.[2] Time does not permit a recitation of the details of these particular investigations. Assuming, however, that systematic observation has some utility for research in the teaching (and learning) of English, I should like to mention now some of the essential characteristics of observation schemes which you could devise for yourselves.

First, the elements of behavior to be observed should be as objective and nonevaluative as possible. We have found that it is difficult, but necessary, to construct a classroom observation schedule in which value laden words are omitted in the categories of behaviors. Thus, if you set up a category of observing teacher behavior such as, "shows good command of language," the interpretation of the word "good" leads to a severe lack of agreement among different observers. In addition, the word "good" promotes a "halo" factor in the observations. It would be better to say, "used an expression not found in so-and-so's list of 20,000 common English words." The categories should be as specific as possible and still be relevant to the objective of describing meaningful events. The weights to be assigned to particular behaviors should, however, emerge from an analysis of the empirical data. The foregoing should not be taken as implying that the observer's function is purely mechanical. On the contrary, the observer's judgment as to how a particular behavior should be coded is crucial and usually calls for great skill.

---

[2] D. M. Medley and H. E. Mitzel, "Measuring Classroom Behavior by Systematic Observation," in N. L. Gage (ed.), Handbook of Research on Teaching (Chicago: Rand-McNally Company, 1963), pp. 247-328.

Second, it is patently impossible to record everything that occurs in a classroom during even a short time span, so the observation schedule should be an abstraction of the relevant behaviors and events from the total complex. Abstraction is a principal tool of science, and it is the job of the observer to perform this abstractive function by attending to and recording selected aspects of the phenomena under study. One may, as Withall[3] and Smith[4] have done in their separate studies, pay attention only to verbal behavior in the classroom. Other investigators have sampled a wider variety of phenomena, including gestures, body movements, facial expressions, as well as verbal behavior. As an example of extreme abstraction, Harrington tallied only smiling behavior, and his data consisted of the number of smiles per unit of time for each of a group of teachers.[5]

A third characteristic of systematic observation, related to the non-evaluative and abstractive functions previously mentioned, is the coding or categorization element. Two approaches to the construction of observation schemes are in wide use—a category or classificatory system and a sign or symptom system. Both are used in conjunction with timed observation periods. A category system is limited to the observation of one segment or aspect of classroom behavior. After determining a convenient unit of behavior, a finite set of categories is constructed, into one, and only one, of which every observed unit is classified. The essential characteristic of the category system is its exhaustiveness, in that every behavior that meets the criterion is classified in some category. To illustrate, Withall decided to categorize teachers' verbal behavior, with the statement or logical thought being the unit.[6] He devised a seven-category continuum from learner-centered to teacher-centered with one of the categories for neutral or otherwise unclassifiable statements.

[3] J. Withall, "The Development of a Climate Index," Journal of Educational Research, 45 (1951), 93-99; and "Development of a Technique for the Measurement of Socioeconomic Climate in Classrooms," Journal of Experimental Education, 17 (1949), 347-361.

[4] B. O. Smith, "A Study of the Logic of Teaching: A Report on the First Phase of a Five-year Research Project" (Washington: U. S. Office of Education, 1959). (Dittoed)

[5] G. M. Harrington, "Smiling as a Measure of Teacher Effectiveness," Journal of Educational Research, 48 (1955), 715-717.

[6] Withall, op. cit.

The sign or symptom system is approached by listing beforehand a number of specific acts or incidents of behavior which may or may not occur during a period of observation. Choice of behaviors to look for is based upon their presumed relationship to a behavior syndrome. An observation schedule based on signs is nonexhaustive in that the observer does not record all that he perceives; instead, he tallies only the behaviors that fit his particular aggregate of signs. He is required to keep in mind a large number of specifics as he observes functioning classrooms. Both approaches provide valuable information about classroom behaviors, and it is likely that any comprehensive observation schedule will have to use both types of items.

Let us assume that an observation schedule containing a large number of specific classroom behaviors has been constructed and that we want to plan for the reduction of data that will be collected in a number of classrooms. In other words, we want to "dimensionalize," if I may coin a word, the thousands of separate observations into measures for each classroom or each teacher. There are two ways of proceeding from this point; both have advantages and drawbacks. One can, on an *a priori* basis, put separate behaviors together to form a logical or educationally meaningful key. The individual items on such a key usually do not possess enough discriminating power to distinguish between different teachers, but we know from Cronbach's work with test items, that items with moderate loadings on a common factor rapidly build up discriminating power, while specific factors tend to cancel each other out.[7] Our experience has been that three to five logically related items from an observation schedule often build into a scale or dimension with internal consistency reliability of .60 or better. Variables elicited from the raw data by the *a priori* method become the bricks for constructing a smaller number of independent measures. The second procedure for "dimensionalizing" classroom behavior data is to apply a factor analytic approach directly to the raw data and to seek the minimum number of factors that will account for the maximum portion of the variance. The factors from this analysis are then useful as measures for quantitatively describing classroom behavior. The trouble with the *a priori* method of item combination is that it is too dependent upon the intuitive skill and insight of the investigator in seeing what goes with what. And the difficulty with the factor analysis approach is that, by capitalizing upon fortuitous events and specific variance in the individual observation items, it puts too many behaviors together and creates too many scales which do not hold up upon a repetition of the study with a different sample. In our most recent study in New York, Medley and I used a combination of these approaches.[8] Our original observation schedule had about 180 distinct items which were reduced by the *a priori* method to 29 variables. A factor analysis of these 29 variables yielded 8 dimensions of classroom behavior which reliably differentiated among a group of student teachers.

---

[7] Lee J. Cronbach, "Coefficient Alpha and the Internal Structure of Tests," *Psychometrika*, 16 (1951), 297-334.

How then does this rather involved approach called systematic observation differ from the more familiar rating scale as a way of studying teaching and learning? As sources of quantitative data about teaching, rating scales suffer from severe shortcomings. The typical rating scale makes impossible demands upon the rater. In effect, it asks him to perceive and store in his memory hundreds of separate behaviors, to classify these behaviors according to their relevancy for some fifteen or twenty separate factors, to weight each behavior appropriately and come up with a composite evaluation for one teacher on each factor without writing down a single intermediate step. It is little wonder that the separate scores derived from most rating scales have been shown to be highly intercorrelated. With systematic observation, we do not suppose that the well-known halo effect is completely eliminated. It is, however, minimized because the judgment made by the observer is a minute one, requiring only a decision as to whether or not a particular behavior or event occurred. The judgment as made is nonevaluative, the weights and values being assigned by clerks and computers in the process of data reduction.

Toward a Theoretical Framework

Although those of us who are engaged in research in classroom behavior know that we have not yet evolved a full-blown theoretical framework into which our findings neatly fit, we are beginning to see some rough outlines of what a theory of teaching-learning might look like. In such a theoretical structure and the research it engenders, the variables derived from systematic classroom observation play a number of roles. First, systematic classroom observation can provide the researcher with an independent criterion measure. For example, it would be possible to study correct pronunciation on the part of classes of ninth grade pupils and to construct a list of commonly mispronounced words. This list could be carried by a classroom observer who could visit ninth grade classrooms and tally instances of mispronunciation by pupils. Class mispronunciation scores could be adjusted for such contingency factors as the number of pupils engaging in class discussion, the amount of time the teacher talked, the number of football players in the class, and so forth. This illustrative structure would have to be predicated upon the idea that correct pronunciation was an important educational goal for ninth graders. The data might or might not tell you anything about the pronunciation of ninth graders in their homes or on the street. We can, however, use an observationally derived variable as an independent criterion and relate to it those dependent measures that concern us. Unfortunately, in most educational research, investigators have assumed that you have to administer a standardized test to pupils in order to obtain a respectable criterion measure. I believe that classroom observation offers a methodology for studying a much wider variety of criteria than has typically been undertaken by means of standardized testing.

---

[8] Medley and Mitzel, *op. cit.*

A second use for observation methodology shows up clearly in research on teacher effectiveness in which it is desired to develop a measure of teaching performance. If we are wise enough in teacher education programs to specify some of the worthwhile behavior patterns that our students ought to adopt, then the measure of the desired behavior becomes a *process* criterion. I distinguish process criteria from product criteria in that product criteria are based upon the assessment of pupil change or growth, whereas process criteria refer to the measurement of desirable teacher and pupil behaviors occurring in the classroom setting. Suppose, for example, we decide that our student teachers should maintain better classroom order, but at the same time they should manifest less hostility toward their pupils. Classroom order and manifest hostility as measures of process can easily be obtained from a systematic observation schedule, and we can use the quantitative values on these dimensions to assess change, to evaluate training methods, or to study the relationship between student teacher personality and student teacher classroom behavior. Eventually we will have to relate these process criteria which seem to be valuable in their own right to appropriate product criteria.

A third use of the variables drawn from systematic observation is as independent control variables in experimentation in classroom methods. My guess is that this use would be of the greatest interest to a subject matter group attending a research design conference. In setting up an experiment in teaching methods, we ordinarily define the separate treatments with great care, making sure that they are as distinct as possible. We carefully instruct the teachers on the employment of method. We assign the pupils by a random process to the various treatments and make provisions for the control of school effects and other environmental factors. We carefully provide for the measurement of pupil achievement at the end of the experiment and analyze our data, firm in the belief that the results will be meaningful. The evidence shows, however, that the typical experiment in methods, which I have described, usually yields no significant differences ascribable to methods or treatment. I believe that the critical factor omitted from this experimental paradigm has been the control measures based on systematic observation of the application of the experimental treatments by teachers and pupils during the course of the study. It should be possible to devise systematic classroom observation procedures which would yield measures of the extent to which teachers and pupils adhered to the specifications of method as planned in the experiment. Measures derived in this way from systematic observation can be used as independent variables in covariance analysis, and to the extent that these observation variables are related to the criteria, the precision of the study will be increased.

Let me tell you the story that leads me to propose this modification in experimentation in educational methods. Recently, in an eastern metropolis, two colleagues and I sallied forth into the public schools to compare individualized instruction in ninth grade algebra with the traditional "lecture-assignment-recitation" method. Nine schools and nine teachers agreed to assist us, each teacher agreeing to teach four sections of algebra, two by a carefully planned experimental method and two by the traditional approach. We conducted a six-week summer workshop for the teachers in advance of the experimental period, being careful to explain the purposes of experimentation and to provide instruction and practice in the application of the two treatments. The experiment got underway as planned, but what we found after a few weeks was that several of the junior high school mathematics teachers in the experiment could not control their own behavior well enough to maintain the "new," or individualized, teaching approach in two of their four sections of algebra. A pupil let the cat out of the bag in one school by telling the principal investigator, "We only individualize on Wednesdays." Wednesday was the regularly scheduled day for a supervisory visit. The point of this story is not that teachers are intentionally trying to "gum up" experiments, but that, by employing some appropriate variables derived from systematic observation techniques, we could have salvaged some of the information from the study. It is naive to expect that teachers in methods experiments will all employ the prescribed treatments to the same degree. The methodology of systematic observation offers us as research workers a source of control data that could substantially improve this type of research.

Replies to Objections

In conclusion, let me react to a few objections that have been raised to the use of the methodology of systematic observation in the study of classroom phenomena. It has been alleged that the cost of placing an observer in experimental classrooms several times a week over a period of months is too great. Some hold that observers in classrooms represent an invasion of the privacy of the teacher and that what goes on in the classroom is privileged communication. A third objection is that placing an observer in the classroom somehow changes the situation so that the obtained measures are distorted.

The objection to the high cost of adding repeated classroom observations to a research study does, of course, have some validity, but in view of our slow progress in educational research over more than a half century we cannot afford to fail to take advantage of any new methodology that offers promise. It used to be that the bulk of educational research stemmed from the short-term, relatively uncoordinated efforts of students on doctoral dissertations. New sources of research support from the federal government and private foundations should remove some of the economic pinch on our research planning. In this connection, I am convinced that we get more new knowledge for our money by doing a sophisticated, intensive study in a small number of classrooms than we do by skimming a minimal amount of data from a large sample.

The objection that observation methodology invades the privacy of the classroom seems to me to have no validity. The notion of the sanctity of the classroom rests on the assumption that the transactions between the teacher and pupil are privileged in the same sense that transactions between doctor and patient, lawyer and client are undisclosed. My guess is that the idea has grown up in the folklore of American public school teaching to protect the teachers from "snoopy" administrators. Our experience as researchers has been that, as soon as teachers understand that our observations have nothing to do with their job security, they do not greatly object to observers either human or electronic.

To the objection that systematic observation distorts the dynamics of the classroom situation there is no satisfactory rejoinder. Because we have no direct way of knowing what happens in the classroom when human or electronic observers are absent, there doesn't seem to be an empirically determinable answer to the question. We get around this objection to some extent by making numerous observations and depending upon a familiarity factor to minimize the distorting effects of our presence on the patterns of classroom interaction. My opinion is that there are so many other relatively uncontrolled factors impinging upon the *observable* behavior in classrooms, it is fruitless to worry about this particular source of bias.

SOME RECENT DEVELOPMENTS IN FACTOR ANALYSIS

Chester W. Harris

The major purpose of this paper is to describe some recent developments in factor analysis that seem to me to be worthy of attention on the part of those who see factor analysis as one useful tool in their research. In order to describe these developments in a meaningful context, I shall cover some old ground that may be quite familiar to you; for this I ask your indulgence. My descriptions will be rather general, without the attention to technical details that would be required in order to satisfy many of my colleagues.

This paper is divided into five sections: Component Analysis, Image Analysis, Scale-Free Factor Analysis, A Note on the Rotation Problem, and Extensions to Multiple Modes. The first three topics were selected to emphasize distinctions among models of factor analysis, in the hope that such distinctions will help provide research workers with better bases for choosing among possible models. Not all the methods or procedures that you may associate with factor analysis are treated in these first three topics; by omitting them I am implying that although they may be interesting or important methods in terms of the history of factor analysis, they appear to have no substantial place today. The book by Harry Harman called *Modern Factor Analysis* is an important historical document, but much of it is quite out of date as a guide to present practice even though its copyright date is 1960.[1]

Let me first acknowledge an indebtedness. For several years a small group of persons interested in the problems of factor analysis has been meeting periodically at the University of Illinois, with Ledyard R Tucker as chairman. This group, whose membership changes at least slightly from time to time, has been discussing questions of what models for factor analysis might exist, how they might differ, and what criteria one might employ in choosing among them. From participating in this group I have learned much, and I am sure that some of what I say here is a report of the views of others in that group, screened—or possibly distorted—by my own perceptions and values. The members also add to my knowledge by bringing "new" ideas to these meetings for criticism and discussion. For example, I shall mention below "alpha factor analysis," developed by Henry F. Kaiser and John Caffrey. The first description of this "new" method was presented in a meeting more than a year ago and thus is well-known to the members of the group even though the paper describing it has yet to be published in *Psychometrika*. Thus any participation in this group not only colors my remarks but also enables me to speak of some developments that are not yet formally documented in the standard journals.

---

[1] Harry Harman, Modern Factor Analysis (Chicago: University of Chicago Press, 1960).

Now let me turn to my five topics.

Component Analysis

It is well known that, given a set of observations on several variables, the data may be reduced to a correlation matrix with units in the diagonals. This correlation matrix may be viewed as a numerical representation of the configuration of the variables *in the variable space*. The dimensionality of this variable space is given by the rank of this correlation matrix, and in practice it usually is equal to the number of variables. This simply says that if we wish a complete representation of the initial set of observations (rescaled and deviated to put them in standard form, with means of zero and equal variances), we nearly always must have as many "factors" as variables. Let us call these "components," rather than "factors." Suppose we wish this complete representation. There is literally no end to the number of ways in which it might be developed. One can take any set of weights which he wishes to dream up and use them to extract a first component from this correlation matrix, which I shall call R. Then, having extracted this first component, one can dream up another set of weights (so long as they differ from the first at least slightly) and extract a second component. And so on. This is an indeterminacy similar to the indeterminacy associated with the rotation problem.

One method of resolving this indeterminacy is to specify a computing algorithm that is to be followed. One can then give the solution consisting of the complete set of components a name descriptive of this computing algorithm. Thus we could develop the complete set of components of R by a centroid algorithm; this would not, however, be the "centroid solution" with which Thurstone is identified, since his work, almost without exception, was done using another model. For modern computing machines, the principal axis transformation is a "natural"; consequently, today, a component analysis—either complete or incomplete—often is achieved by the principal axis transformation, and the solution described as a principal components analysis of R.

One of the chief features of a complete component analysis is that the data that generate the matrix being analyzed can be expressed as a product of two matrices: The first describes the variables in terms of components, and the second describes, say, the individuals on whom the observations were based in terms of components. This first matrix is what is often called a "factor matrix" and the second is what is often called a "matrix of factor scores, or measurements." Some computing algorithms such as the principal axis transformation give what is known as an orthogonal solution. This simply means that the *second* matrix consists of component scores, or measurements, that are uncorrelated (in the sample).

If we ignore the varieties of component analysis that arise from adopting different computing algorithms, we find that there are nevertheless varieties of two additional kinds. Ignoring the varieties associated with the way in which the components are determined is sensible on two grounds. One is that the complete component analysis done by any one method, or computing algorithm, is an orthogonal transformation, or "rotation," of the complete component analysis done by another method; the second is that the principal component analysis, which is well adapted to modern computing machines and has the property of yielding a first component of maximum variance, a second with the next largest variance, etc., has essentially won out in this area. For such an analysis, the matrix of component scores or measurements is rather easy to compute, and the component scores are uniquely defined even though one "discards" the components with the small variances. It is this distinction between a complete and an incomplete component analysis that yields one set of varieties of component analysis, and one reason for preferring a principal components solution is that the partial or incomplete solution yields, for the components that are solved for, the same component measurements that would be secured from a complete solution.

Another dimension along which varieties of component analysis lie is generated by modifying the scale of the original variables. When we compute a correlation matrix, we are arbitrarily changing the scale of the original variables; regardless of the variance of these original variables, each variable is transformed so that all have equal variances. Thus we could generate a variance-covariance matrix from the original variables, retaining their initial scale, in place of the correlation matrix which assigns equal variances to the variables. Now a principal components analysis of the variance-covariance matrix will *not* yield a solution that is related in any simple fashion to the principal components analysis of the correlation matrix. We can generalize this by stating that the principal components solutions of all arbitrary rescalings of the variance-covariance matrix differ from each other in this particular sense. In other words, component analysis is *not* what I shall call a "scale-free" analysis.

The complete principal component analysis nearly always yields as many components as variables. If all these are retained, one has achieved a derived set of variables that are uncorrelated in the sample; for some purposes, it is convenient to transform the original variables into an uncorrelated set. One must recognize, however, that had he begun with a different set of scale values—e.g., the variance-covariance matrix rather than R—he would have secured a different set of uncorrelated variables. The incomplete principal component analysis, in which the components with small variance are discarded, also yields an uncorrelated set of derived variables, but these too differ, depending upon the scale values that are used. Both the complete and the incomplete principal component analysis have the good feature of yielding computable (as opposed to estimated) component measurements, and these are invariant or unique regardless of how many of the components are extracted from a given matrix. Essentially, principal component analysis is a descriptive procedure yielding uncorrelated derived variables that are dependent upon the scales used in the analysis. I prefer not to call this factor analysis. Thus, in my terms, one who extracts the principal components of R corresponding to the roots or eigenvalues greater than unity is not performing a factor analysis. I do not necessarily urge that he abandon this practice; I merely wish that he would label what he has done precisely.

Image Analysis

I shall refer you to three papers. These are by Guttman,[2] who originated image theory, Harris,[3] and Kaiser,[4] in alphabetical and chronological order. An intuitive notion of an image matrix can be developed by thinking of the vectors representing a set of variables as arranged in a space (usually) of as many dimensions as we have variables. The angular separation of the vectors is a function of the correlation between variables, and is given precisely by equating the cosine of this angle to the product-moment correlation coefficient. This angular separation is not dependent upon the scale of the variables; thus we may think of either the variance-covariance matrix or the correlation matrix as represented in this fashion, the only difference being in the lengths of the vectors. For the representation of the correlation matrix, all the vectors are of equal (unit) length.

Having visualized this configuration of vectors, one next imagines that each vector is projected into the space defined by the remaining set of vectors and that we make up a variance-covariance matrix representing these projected vectors. In general, these projected vectors will each now have been shortened somewhat; the length of each projected vector will be equal to the multiple correlation of that variable with the remaining variables. Furthermore, the angular separations of any two of these *projected* vectors will in general differ from the angular separations of the corresponding two original vectors or variables. All these projected vectors will still lie in the variable space, and thus an analysis of these projected vectors, or images, will yield both image factors and image factor measurements that can be computed. I am merely saying that a computable linear transformation of the original data yields a matrix that can be expressed precisely (within rounding error) as the product of the complete image factor matrix and the matrix of image factor measurements.

---

[2] L. Guttman, "Image Theory for the Structure of Quantitative Variates," **Psychometrika**, 18 (1953), 277-296.

[3] C. W. Harris, "Some Rao-Guttman Relationships," **Psychometrika**, 27 (1962), 247-263.

[4] H. F. Kaiser, "Image Analysis," in Chester W. Harris (ed.), **Problems in Measuring Change** (Madison: University of Wisconsin Press, 1963).

Since the image system is scale-free, it is convenient to develop the image matrix from R, the correlation matrix, using the following definition:

$$G = R + S^2R^{-1}S^2 - 2S^2,$$

in which G is the image variance-covariance matrix, R is the correlation matrix with units in the diagonal, and S^2 is a diagonal matrix for which the diagonal entries, s_j^2, are the reciprocals of the diagonal elements of R^{-1}. The square of the multiple correlation of any variable, j, with the remaining set of variables is given by $1 - s_j^2$. The image variance-covariance matrix, G, has these squared multiple correlations in the diagonals and has the important property of not having any negative roots, or eigenvalues. This means that G may be factored completely, usually giving as many image factors as variables, without one's encountering imaginary factors.

I believe that the complete image analysis may prove to be very useful as a guide to test development. My speculation—and it is primarily that—is that the complete set of image factors when rotated orthogonally by the normal varimax procedure will give us a good index of the number of common factors that *should* be regarded as belonging to the domain from which the variables were drawn. Usually such a rotated image factor matrix has three types of columns: those with only trivial entries, and thus those that can be ignored for practical purposes; those that look like common factors, i.e., those with many trivial entries plus a few substantial ones; and those that look like specific factors, i.e., those with only one substantial entry and the rest trivial. It is my hunch that each of these columns that looks like a specific factor may be regarded as evidence that another common factor exists in the domain of variables that have been sampled, but that only one variable that is "loaded" on that common factor happens to have been included. I have tried this with one or two small matrices that are well understood, discarding all but one of the variables that "loads" on a given common factor, and found that the complete image analysis retains a psuedo-specific factor for that variable. Much further investigation needs to be done, but this seems to be a promising development.

An incomplete image analysis has also intrigued me, and I have worked out some relationships between it and what has come to be a fairly common practice, that of factoring the correlation matrix with squared multiple correlations in the diagonals. In the notation used here, this practice is that of factoring $R - S^2$, which is at least superficially different from factoring G, the image variance-covariance matrix. Now $R - S^2$ always has some negative roots, and so it always has some imaginary factors. Guttman has developed a best lower bound for the number of common factors, on the assumption that one has in hand a sample of variables for a population of individuals.[5] This lower bound is a psychometrically oriented rule, rather than a statistically oriented one, in the sense of sampling persons. Let us call this lower bound for the number of common factors m. I have found that when they are extracted in a certain fashion, the first m factors of $R - S^2$, none of which is imaginary, are related in a very simple way to the first m factors of G, the image variance-covariance matrix. Further, we now know that the least squares estimates of the factor measurements for these first m factors of $R - S^2$, when these are extracted in my fashion, are identical with the first m image factor measurements (which can be computed, rather than having to be estimated). By putting together the first m factors of $R - S^2$ and the related image factor measurements, we achieve an approximation of a particular kind to the image variance-covariance matrix, G, itself. Thus I have satisfied myself that factoring $R - S^2$ is merely a back door to an incomplete image analysis. I do not object to a visitor's using the back door; I do however wonder about his spatial orientation or his eyesight when he mistakes this for the entrance to the temple of factor analysis.

Component analysis, either complete or incomplete, and image analysis, either complete or incomplete, have in common the property that the analysis is being carried out in the space of the variables, and thus the associated component or image factor measurements can be computed precisely, rather than having to be estimated. Both procedures are important in their own right and deserve the distinctive names that I believe should be used for them. Neither belongs to what is often called the Spearman-Thurstone tradition of factor analysis, since this tradition is clearly one of postulating a common factor space that is distinct from and not embedded in the variable space. As a consequence, no factoring method that adheres faithfully to this tradition—which emphasizes communalities—can produce unique and computable factor measurements. This is merely a fact and not necessarily a fatal flaw. It is an important fact, however, for those who see a set of measurements for the individuals as a major outcome of their analysis.

---

[5] L. Guttman, "Some Necessary Conditions for Common Factor Analysis," Psychometrika, 19 (1954), 149-162.

If you don't like factor analysis—and I know persons who honestly hold such an antipathy—I recommend to you either the well-known component analysis or the newer and less well-known image analysis which differs at least in being scale-free. However, if you are neutral, or possibly favorably inclined toward factor analysis, consider the following methods.

Scale-free Factor Analysis

I have, so far, suggested that neither the factoring of R nor the factoring of $R - S^2$ is an example of factor analysis. Let me now be more positive. I begin by stating a preference for scale-free solutions. I believe that, in time, only such solutions will be regarded as desirable. This will mean that popular procedures like the centroid will no longer be used, except possibly as means to other solutions.

Original developments by Lawley,[6] Rao,[7] and Anderson and Rubin[8] have given us a class of maximum likelihood solutions, each of which is scale-free. The general character of this class of solutions can be suggested by referring to Rao's canonical factor analysis. Rao proposed that the solution be accomplished by finding the canonical correlations between the observed variables and the set of hypothetical factor measurements. A critical unknown in this system is the set of unique variances, often designated as the diagonal matrix U^2. Thus trial values for U^2 are taken, under the restriction of a stated hypothesis of a given number of common factors, an iteration to achieve stable values for the unique variances under this restriction is made, and a statistical test is conducted. If this test, which depends in part upon the size of the sample of persons, leads to a rejection of the hypothesis of the stated number of common factors, then the process is begun again, postulating one additional factor. This procedure is continued until the number of common factors has been determined in accordance with the statistical test.

---

[6] D. N. Lawley, "The Estimation of Factor Loadings by the Method of Maximum Likelihood," **Proceedings of the Royal Society of Edinburgh. Vol. 60.**

[7] C. R. Rao, "Estimation and Tests of Significance in Factor Analysis," **Psychometrika,** 20 (1950), 93-111.

[8] T. W. Anderson and Herman Rubin, "Statistical Inference in Factor Analysis," **Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability. Vol. V.** (Berkeley: University of California Press, 1956).

Obviously it is not feasible to carry out such a procedure on desk calculators even for relatively small sets of data. We have come to the point where computing facilities that will make literally millions of calculations in a relatively short time are indispensable for research using any of the models that I am describing in this paper. Like component analysis and image analysis, my preferred varieties of factor analysis are geared to the machine age. Merely having a machine, however, does not solve all the problems. For example, there is more than one computing algorithm that might be employed in developing a maximum likelihood solution, and we have not yet finally determined which of them is the best. We do suspect that Lawley's first method doesn't work very well, but the relative merits of the others—and of new ones that might be developed—are not completely assessed. The interaction of the machine and the model has become an area of investigation. Today one needs an excellent numerical analyst as well as a passable factor analyst to bring off a project of any size.

The major feature of the maximum likelihood class of solutions is its inferential character in the standard statistical sense. The number of common factors is determined by a test of significance, and thus one *may* be at the mercy of the power of this test. If so—and I must emphasize the *if*—then using the same variables but using a larger sample of individuals may lead to identifying more common factors. For some, this is an unpleasant possibility; surely the number of common factors should not be a function of the number of individuals tested, but of the variables employed. Two things should be said. For one, how much difference this unpleasant possibility would actually make in the final rotated solution has not yet been assessed to my knowledge; it seems quite possible that an additional factor or two, each of quite small variance, might not disturb the major features of the final rotated solution. Second, a very good point is that this concern about the number of common factors properly being related to the character of the variables rather than to the number of persons tested has given rise to at least the beginnings of what one might call psychometric, rather than statistical, inference. Guttman's image theory and his lower bound for the number of common factors reflect this point of view.

Following Guttman's lead, I recently proposed a variant of Rao's solution that could not possibly satisfy any first-class statistician.[9]

---

[9] Chester W. Harris, "Canonical Factor Models for the Description of Change," in Chester W. Harris (ed.), Problems in Measuring Change (Madison: University of Wisconsin Press, 1963).

In this I abandoned the statistical test in favor of retaining the minimum number of common factors that is specified by Guttman's best lower bound to this number; yet I retained the form of Rao's solution and thus have a scale-free set of common factors. It generally turns out to be a high-communality, many-common-factors solution, which will displease some, I am sure. It does this by retaining all the common factors for which, in the sample, there are positive canonical correlations between the set of variables and the set of hypothetical factor measurements. As this statement suggests, the factor measurements cannot be calculated uniquely but must be estimated; this is true for the class of maximum likelihood solutions generally.

This variant of Rao by way of Harris rescales the variables in the metric of the unique portions of the data, analyzes the rescaled data, and then transforms the results back into the original metric of the variables if that is desired, or into any other metric one wants, such as the equal-variance metric associated with standard scores. Thus the rows of the factor matrix itself can be weighted to correspond to hypothetical or desired patterns of weights for the variables. If the normal varimax rotation procedure is used on the factors to derive a rotated solution, these rotated factors can also be weighted in this fashion; practically all other rotation procedures would destroy this correspondence.

This notion of rescaling the variables in the metric of the unique portions of the data is relatively old. A different approach has been taken by Kaiser and Caffrey in their alpha factor analysis.[10] They propose that the variables be rescaled in the metric of the common, rather than the unique, portions of the data. Once again, the rescaling is done, the analysis carried out, and then the results are transformed back into the original metric or any variant of it that is desired. It also is true for alpha factor analysis that normal varimax rotations will not destroy this scale-free or scale-separability feature.

Perhaps the most interesting feature of alpha factor analysis is the rule for the number of common factors. Recall that both for the incomplete image analysis I have described and for the variant of Rao that I developed, Guttman's lower bound, *m*, is taken as the number of common factors. This gives the common factors for which nonimaginary canonical correlations between factors and variables exist for the finite set of individuals who were tested. Kaiser and Caffrey examined the reliability or generalizability of factors and developed their rule so as to determine from the variables in hand those factors that can be generalized to the universe of variables, or universe of the content. This is a psychometrically oriented rule, guiding on inference from a sample of variables to a universe of variables. I am fairly certain that in most cases the Kaiser-Caffrey rule gives fewer common factors than does the rule associated with Guttman's *m*. Thus, interesting theoretical questions are raised. In addition, empirical work that will demonstrate the magnitude of the differences in the final rotated solutions that are derived *from* these two points of view is clearly called for.

---

[10] H. F. Kaiser and John Caffrey, "Alpha Factor Analysis," Psychometrika (in press).

I hope that my discussion of these first three topics has communicated to you my feeling that many interesting developments in connection with models for the analysis of the interrelationships of variables have occurred recently and are still occurring. There is vitality here. Now let me suggest a similar vitality in connection with the rotation problem.

A Note on Rotation

Harman's book has two chapters on analytic—or machine—methods of rotation.[11] Machine methods have advantages: they are economical and they are objective. But they also have the disadvantage of getting the wrong answer at times. Harman judges, and I agree with him, that the problem of analytic orthogonal rotation has practically been solved, even though we may still expect further minor modifications. The problem of analytic oblique rotation seems to him and to me to be at a different stage of development. What one would like is a new view of this problem that might simplify the procedures but at the same time show how to get the right answers, when this is possible, and otherwise show why a particular set of right answers cannot be secured.

Both Kaiser and I have, at times past, shown a certain amount of irreverence for what others regard as the eternal verities of factor analysis. It probably is fitting therefore that he and I should collaborate in attacking this oblique rotation problem in a new way. This may sound startling, but what we have done is to throw away the conventional rules and develop procedures for deriving rotated oblique solutions by making *orthogonal* transformations; in particular, we are using the normal varimax procedure to make these transformations since he likes it very much and I concede its merits. The secret of our procedure is that we don't rotate the original factor solution; instead we rescale it and, by rotating the rescaled original factor solution, derive an oblique solution. You are already aware that both Kaiser and I find rescaling a powerful tool; it is not surprising that we carry this idea over to the rotation problem.

Our investigation has given us a system for classifying all the possible derived solutions for any given original factor solution. From this investigation we also conclude that there are two mutually exclusive classes of oblique solutions, rather than only one. Within these classes we are identifying particular types of solutions that have interesting features. For example, we can now provide an unambiguous and direct method for testing one's hunch that the common factor portions of a particular set of variables can be expressed as a set of independent (but possibly correlated) clusters. Perhaps you will forgive the enthusiasm with which I report that we can also solve the famous box problem of Thurstone; those of you who have tried to do this with analytic rotation procedures know that this is at least a mild achievement. An initial paper was presented to the group that meets at Illinois and survived the scrutiny of persons like Ledyard Tucker and Paul Horst. We believe that we are in business; soon we shall try to publish a report.

---

[11] Harman, op. cit.

Much more might be said about the rotation problem; but at the moment I think that every method is headed toward oblivion except ours, and so I close this section on that immodest note.

Extension to Multiple Modes

Let me close this paper by discussing very briefly another line of development that I believe will become quite prominent in the 1960's. We can illustrate the need for this development by considering the technique known as the semantic differential in which individuals judge or rate concepts or terms on a set of scales. This is a three-mode system rather than a two-mode system; the three modes are *entities* (the individuals, e.g.), *attributes* (the scales), and *referents* (the concepts or terms). The models I have discussed provide for the treatment of data arising from only two modes: entities and attributes, generally. What then is to be done if we want to do a factor analysis of data gathered by the semantic differential technique? I know what people have done. They have collapsed one of the modes, and, for example, factored data consisting of group means for each scale and each concept.

Possibilities for extending the notions of component analysis and of factor analysis to multiple modes are now being developed. So that you may refer to something definite on this topic, I cite the recent papers by Horst[12] and Tucker.[13] I suspect that you will find these papers challenging. The potentiality of the developments outlined in these papers seems to me to be very great. Three-mode sets of data are very common: raters, ratees, and scales; or individuals, variables, and occasions, as in longitudinal studies; etc. Four- and five-mode systems are not impossible

[12] P. Horst, "Multivariate Models for Evaluating Change," in Chester W. Harris (ed.), **Problems in Measuring Change** (Madison: University of Wisconsin Press, 1963).

[13] L. R Tucker, "Implications of Factor Analysis of Three-way Matrices for Measurement of Change," in Chester W. Harris (ed.), **Problems in Measuring Change** (Madison: University of Wisconsin Press, 1963).

to conceive. The development is badly needed in order to handle these multiple mode sets of data more completely and with greater elegance. A number of problems will arise, some of them paralleling the problems we have wrestled with in the conventional two-mode case. At the moment, a components-type approach to multiple modes seems to be farthest along; extension of the notion of communality or communalities to multiple mode analysis will generate many interesting questions. But in time we should be able to handle in a better fashion sets of data that we now can treat only in segments or slices or by collapsing a dimension that is potentially very important.

THE ANALYSIS OF PREFERENCES

E. A. Peel

The Problem

I begin with the trite remark that when one expresses a liking or preference for a picture, poem, novel, or piece of music, we have the result of the impact of a work of art on a person. A work of art has recognizable features, some objective in the sense that their recognition demands no agreement among experts, some so much more intangible and variable that a consensus of opinion is needed, others so difficult to pin down that they lie in the eye of the individual beholder. On his side the person is even more complex, and his reaction to a work of art involves feeling, perception, intelligence, experience, and background. The impact of these two sets of circumstances brings about the expression of opinion. Both sets of conditions operating together in some way determine what the preference will be. Thus differences in appreciation have been traced to personality differences such as Bullough's types (associative, objective, character),[1] extraversion, introversion, radicalism–conservatism and youth–age (here is an objective quality of a person).

Appreciators of poetry have been similarly differentiated. Most attention was first given to the second types of analysis of preferences in terms of the personal qualities of the appreciator.

The basic method in these researches was to ask the persons to show their preferences from a set of paintings, poems, or prose pieces and then for the psychologist to use independent methods for assessing the person's psychological or social qualities, as done by Burt,[2] Eysenck,[3] and more recently Burton.[4]

When I first became interested in the aesthetic appreciation of painting (I should say I am a sporadic Sunday painter), nobody seemed to have given serious attention to the alternate approach implied in my opening remarks–that of analyzing the preferences in terms of the qualities of the work of art. Thus I wrote: "Where identification of aesthetic factors has been attempted, it has been with reference to temperamental traits, and the emphasis has been on the person rather than the artistic qualities of the picture."[5]

---

[1] E. Bullough, "The Perceptive Problem in the Aesthetic Appreciation of Single Colours," **British Journal of Psychology, II** (1908), 406-486.

[2] C. Burt, **The Factors of Mind** (London: University of London Press, 1940), Chapter V

[3] H. J. Eysenck, "The General Factor in Aesthetic Judgments," **British Journal of Psychology, XXXI** (1940), 94; "Some Factors in Appreciation of Poetry," **Character and Personality, IX** (1940), 160; and "Type Factors in Aesthetic Judgements," **British Journal of Psychology, XXXI** (1941), 263.

[4] D. L. Burton, "The Relationship of Literary Appreciation in Certain Measurable Factors," **Journal of Educational Psychology, 43** (1952), 436.

The two approaches are complementary. The one seeks to verify whether A likes this picture or poem or joke because, say, he is an extravert feeling type,[6] the other whether A likes it because the picture is naturalistic or the poem expresses conflict, etc.

I concentrated upon the second type of analysis of liking in terms of the artistic qualities of the material and devised a method of analysis which has been used with some success in the appreciation of art, poetry, and prose and in the analysis of assessments of essays.[7]

The Method of Studying Preferences

The method is really quite simple. Suppose we had a test composed of a number of pictures, short stanzas of poetry, jokes, short extracts of prose or essays; we can ask a person to score these creations in accordance with his preferences. This scoring may take the form of placing, say, the pictures in an order of preference or they may be given scores on whatever basis is suitable, e.g., essays on a fifteen-point scale or poems on a five-point scale of liking. This gives the rank order of what I call the *criterion* (of liking).

We may now get experts to look at the pictures or writings in terms of any artistic, literary, or humorous qualities which they may manifest. Thus, supposing I had selected a set of pictures ranging at one extreme from abstract painting through various intermediate phases to the most concrete, giving the illusion of reality, I could ask the experts each to order or score the picture items on the scale of abstraction-realism. I should, of course, define my terms more carefully than I am doing in this paper. Similarly a set of poems could be selected and ordered for simplicity-complexity, or on a scale of restrained feeling–great emotionality. This set of pictures or poems might possess several of these qualities: color, use of concrete allusion, and poetic versus nonpoetic diction.

These orders or scorings of the items I call the *determiners* since to a greater or less degree they may be said to *determine* the *preferences.* I may give the sets to be ordered for liking to groups of people–children, lay adults, professionals—and analyze their preferences by correlation methods in terms of the determiners chosen. I then have correlations between persons' liking, between likings and determiners, and between determiners; an ideal set of determiners is one whose members correlate little with each other. These correlation tables give rise to various interesting methods of analysis, including factor analysis and multivariate analysis described elsewhere.[8]

---

[5] E. A. Peel, "On Identifying Aesthetic Types," **British Journal of Psychology**, XXXV (1945), 61.

[6] H. Read, **Education Through Art** (New York: Pantheon Books, 1958).

[7] Soon after this time Osgood's semantic differential appeared, by which one may also analyze a preference in terms of the qualities in the work of art, but I was not able to see its application to passages from novels until recently when one of my research students, Harpin, successfully used it in a study of appreciation. He applied a set of twenty scales including measures of **warmth, sincerity, vividness, depth,** and **clarity,** etc., to extracts from the writings of twelve novelists including Forster, de la Roche, Pater, Defoe, and Farnol. See W. S. Harpin, "The Measurement and Evaluation of Prose Appreciation with Especial Reference to the Novel Using the Semantic Differential" (Unpublished M. Ed. thesis, University of Birmingham, 1962).

Some Analysis of Preferences

I shall discuss briefly the results of three experiments, one undertaken by me in visual art, one on poems by Britton,[9] and one on the marking of children's compositions by Percival.[10]

In my test I used a set of thirty-one postcard-size colored reproductions of landscape painting including work by artists from the Middle Ages to modern times. The theme was confined to landscapes to restrict irrelevant influences introduced by diverse subject matter and content.

Three artistic qualities were chosen: *detailed naturalism,* resulting from what has been called consecutive vision, seen in Frith's "Derby Day"; *composition,* the unified, satisfying arrangement of the elements of a picture; *atmospheric light,* distinguishing bright light pictures from sombre landscapes.

The results of the regression analysis were as follows:

		Determiner Weights			
	Person	Detailed Naturalism	Composition	Atmospheric Light	Maximum Correlation
Artists	A	–49	42	–33	70
	B	–42	28	–31	56
	C	–28	51	–09	61
	D	–01	67	–12	76
Non-artists	E	60	14	–21	68
	F	70	12	–11	74
	G	47	07	02	46
	H	26	23	–24	44
	J	–18	20	04	30
	Q	14	36	–20	43
	L	–46	–06	–26	47
	M	46	31	–01	52

---

[8] E. A. Peel, "Analyse Factorielle de Correlations entre Personnes," **L'Analyse Factorielle et ses Applications** (Paris: Centre National de la Recherche Scientifique, 1955), pp. 375-392; "A New Method for Analysing Aesthetic Preferences: Some Theoretical Considerations," **Psychometrika,** 11 (1946), 129; and Peel, op. cit. (Note 5).

[9] J. N. Britton, "A Study of Responses to Poetry on the Part of Adolescents and Adults" (London: British Association, York, 1959).

[10] E. Percival, "A Quality Scale of English Compositions for an 11 Plus Age Group" (Unpublished M. A. [Education] thesis, University of Birmingham, 1958).

We see the artists generally disliked the naturalistic paintings and preferred those showing good compositions. These results are confirmed by the factor analysis which produced the following rotated loadings:

	Rotated Loading			
	I		II	
Adult	+	−	+	−
A		54	52	
B		46	29	
C		32	53	
D		17	75	
E	66		05	
F	78		08	
G	52		01	
H	42		15	
J		08	19	
K	25		58	
L		43		01
M	59		37	
Detailed Naturalism	[90]		00	
Composition		12	[85]	
Atmospheric Light		27		27

Without going into more statistical and psychometric details, it is clear that the determiners relate significantly to the liking choices of the twelve persons investigated.

Now let us look at material of more interest to English specialists. Britton applied the technique to an analysis of the preference ratings of twenty-four subjects for thirty-one short modern poems.[11] He finally chose three determiners. First he had *simple* versus *complex* poetry. Then he pointed out that "the poems varied in the degree to which the writer stressed *conflict* between their elements (love and war, life and death, old age and youth) and the degree to which he stressed a resolution."

As a poem revealing conflict, he selected "The End of a Leave" by Roy Fuller.

> Out of the damp black light,
> The noise of locomotives,
> A thousand whispering,
> Sharp-nailed, sinewed, slight
> I meet that alien thing,
> Your hand, with all its motives.
>
> Far from the roof of night
> And iron these encounter;
> In the gigantic hall
> As the severing light
> Menaces, human, small
> These hands exchange their counters.
>
> Suddenly our relation
> Is terrifyingly simple
> Against our wretched times,
> Like a hand which mimes
> Love in this anguished station
> Against a whole world's pull.[12]

---

[11] Britton, op. cit.

[12] Roy Fuller, Collected Poems (Chester Springs, Pa.: Dufour Editions, 1962, and London: André Deutsch Ltd.). Reprinted by permission of Roy Fuller and Dufour Editions.

And as an example of a resolving poem in which the outcome is foreseen and accepted, he used Robert Graves' "Despite and Still."

Have you not read
The words in my head,
And I made part
Of your own heart?
We have been such as draw
The losing straw—
You of your gentleness,
I of my rashness,
Both of despair—
Yet still might share
This happy will:
To love despite and still.
Never let us deny
The thing's necessity,
But, O, refuse
To choose
Where chance may seem to give
Loves in alternative.[13]

To his team of seven expert judges he gave this wording:

"At the top of the list: poems which cherish the feelings, hopes, etc., of the individual in opposition to whatever may be in conflict with them. At the bottom of the list: poems which show a willingness to accept the loss of individuality in something more powerful or more significant than the individual man, so suggesting a resolution of the conflict."

---

[13] Robert Graves, Collected Poems 1955 (Garden City: Doubleday and Company, 1955), p. 187, and Collected Poems 1959 (London: Cassell and Company Ltd., 1959), p. 193. Reprinted by permission of International Authors N. V., holders of all copyrights.

3

For his third determiner he chose the scale represented by the extremes: *incidence of everyday words and structures versus poetic words and structures.* These are exemplified in the following:

(Louis MacNeice, "Entirely")

If we could get the *hang* of it entirely
It would take too long;
All we know is the *splash* of words in passing
And falling twigs of song,
And when we try to eavesdrop on the great
Presences it is rarely
That by a *stroke of luck* we can appropriate
Even a phrase entirely.[14]

(Humbert Wolfe, "Trojan Woman")

So *let* the tears resume to tell us how
There is nothing left of any words she spoke
But *blossom scattered* with the breaking bough,
The *cadence broken* when her own heart broke.
Nor read of *fabulous* women *loved and fair*
In Troy or Babylon, whose *sorrow lingers*
Upon the bright and *adventitious air*
Of Beauty nor their own, but their gold singer's.[15]

This third determiner, *everyday versus poetic diction,* had greater influence in the choice of the twelve nonexperts who included science graduates, an economist, and a bank clerk than did the other determiners. The second determiner, *conflict versus resolution,* was of greater importance in determining the choices of the twelve experts (English honors graduates who were also teachers of English especially interested in poetry).

	Everyday-Poetic Diction	Simple-Complex	Conflict-Resolution
Factor	[.83]	00	.44
	.13	.81	.27
	.04	.20	.75
Variance	9½%	13½%	13%
Nonexperts	12½%	12%	8%
Experts	7%	15%	18%

[14] Louis MacNeice, Collected Poems, 1925-1948 (New York: Oxford University Press, 1963, and London: Faber and Faber Ltd.). Reprinted by permission of Faber and Faber and Oxford University Press.

[15] Humbert Wolfe, "Trojan Woman." Copyright owned by Miss Ann Wolfe. Passage reprinted by her permission.

To quote Britton:

> I suggest then, in simple terms, that a preference for a certain kind of language—either conventionally poetic or conversational language—influenced these subjects in their judgment of these poems, and that it influenced the nonexperts nearly twice as strongly as it did the experts (on average). Secondly that a general attitude to life exhibited in the poems—either on the one hand a zest for the conflict of the individual against opposing forces, or on the other hand a sense of the individual's ability to find satisfaction in a world of conflicting forces—was also an important influence upon these subjects' response to these poems: and that this consideration affected the experts well over twice as much as it did the nonexperts.[16]

The last research I want to mention shows how the technique can be applied to analyze the basis of overall impression methods of marking children's compositions. Percival asked judges to evaluate the quality of the compositions of some hundreds of children on the same topic, "How the Twins Found the Treasure"—started off for them by means of a picture strip.[17] He then analyzed the composition independently to obtain the following determiners: number of words, average length of sentence, number of coordinating conjunctions, different kinds of punctuation, subordinating conjunctions, effective language, accuracy of composition, accuracy of spelling.

---

[16] Britton, op. cit.
[17] Percival, op. cit.

He obtained the following correlations between the criterion *quality* and various teams of determiners equally weighted.

CORRELATION OF "GENERAL COMPOSITION QUALITY" WITH A NUMBER OF COMPLEX DETERMINERS

Criteria	*Correlation*	
	Boys	*Girls*
1. Intelligence Quotient + Education Quotient + Number of Words	.84	.83
2. Education Quotient + Number of Words + Accuracy of Spelling + Accuracy in Composition + Punctuation	.86	.83
3. Education Quotient + Number of Words + Accuracy of Spelling + Accuracy in Composition + Punctuation + Effective Language + Sentence Length	.85	.83
4. Effective Language + Sentence Length + Coordination + Subordination	.67	.67
5. Coordination + Sentence Length	.57	.49
6. Effective Language + Sentence Length	.64	.62
7. Effective Language + Subordinating Conjunctions	.64	.68
8. Coordinating Conjunctions + Effective Language	.70	.61
9. Coordinating Conjunctions + Subordinating Conjunctions	.58	.58
10. Accuracy in Composition + Accuracy in Spelling	.75	.71
11. Number of Words + Accuracy in Composition + Accuracy of Spelling	.83	.79
12. Education Quotient + Number of Words + Punctuation	.83	.76
13. Education Quotient + Number of Words	.81	.79
14. Intelligence Quotient + Education Quotient	.81	.83
15. Number of Words + Effective Language + Accuracy in Composition + Accuracy of Spelling	.93	.81

When we calculated the weights leading to maximum prediction, we found the most predictive determiner was the number of words in a composition!

Implications for the Analysis of Literary Preferences

Although the research technique started as a method of analyzing aesthetic preferences in the field of painting and visual art, it seemed that the method or some modification of it might be applicable in the literary field; this turned out to be the case. But much more research is probably required in order to make the technique more effective. One big difference between the two kinds of appreciation is obvious at the start. A painting evokes a more or less instantaneous attitude, whereas, as Britton remarks, poetry and prose have to sink in. The method is limited to the appreciation of short passages.

One of my research students has also applied it with moderate success to the analysis of people's reaction to jokes. Here it is not easy to obtain continuous determiners running through all the jokes. It is better to use a dichotomous distribution as, for instance, in aggressive versus sex jokes. We have statistical devices to cope with such a situation.

Both preferences and determiners depend to some extent upon the selection of pictures and literary extracts chosen, but there is no real weakness in this. One may have formed a hypothesis which can only be tested by a selected sample of materials. On the other hand, one might attempt to draw a random sample in order to examine first the extent of general agreement among judges concerning a random sample.

The method outlined or some near variation of it could be used in any field of human evaluation where one would wish to discover something about the basis of the evaluation.

IV. SUMMARIES AND CONCLUDING STATEMENTS

The brief reports in this section illustrate once again that different groups react differently to the same papers or ideas and that they go about their tasks in very different ways. Each group enjoyed rigorous discussions, not here recorded, and worked on (if they did not always complete) a typical design for a research study in its particular area of study. The following reports include two preliminary designs executed in some detail and many specific suggestions for other patterns and procedures for research in English teaching. Perhaps of even greater interest is a considerable collection of miscellaneous ideas which apparently appealed to the recorders. The editors hope that many of them will also attract the reader and provide possible guides for the prospective research worker.

Mr. Ianni, Director of the Cooperative Research Branch of the United States Office of Education, reviews some of the history of Project English, an important part of the research activity supported by the Branch, and indicates some of the future possibilities in expanding research and development projects in the teaching of English.

The final statement appraises some of the results of the conference and offers suggestions for future activities.

REPORTS OF STUDY GROUPS

Research in Teaching Language[1]

Finding a definition of language that all would accept was the first problem that the study group on Teaching Language encountered. Essentially the issue was whether language includes written expression. The issue was not resolved to the satisfaction of all participants, even after the several sessions, though most were willing to accept language as speech—a system that can be described at the phonological and grammatical levels—and to regard spelling, composition, and semantics as being the concern of other study groups.

A second problem, related to the first, occurred in attempting to define the objectives for teaching language: Should the effects of language teaching be considered in terms of knowledge about language—language as subject matter—or in terms of oral and written composition—language as used?

The lack of agreement concerning definitions and objectives was not without relevance to the assigned topic, Research Design. One implication is that an individual's point of view determines the nature of the procedures and instruments used in research. For example, if the objectives of language teaching are considered to be intrinsic to the subject matter itself, then measurement of the effects of this teaching would be the determination of the amount of knowledge the students have of the subject matter presented. If the objectives are extrinsic, measuring instruments should be applicable beyond the bounds of the particular teaching and should determine the effects of the teaching upon factors relevant to the objectives. A second implication concerns the nature of the comparisons to be made in certain experiments. For intrinsic objectives it is perhaps specious to attempt to define "control groups." If an objective, for example, is the mastery of the principles of a transformational grammar, it would not be "good" design to contrast the effects of teaching a course based upon transformational grammar with the effects resulting from a course that had some other base. On the other hand, extrinsic objectives might be applicable to many kinds of instruction, and comparison of contrasting procedures and content with respect to such extrinsic criteria would be appropriate.

---

[1] **Group Roster: Leader,** Garlie Forehand, Carnegie Institute of Technology; **Recorder,** Walter Petty, Sacramento State College; Russell Cosper, Purdue University; Naomi Chase, University of Minnesota: Eldonna Evertts, University of Nebraska; Morris Finder, Western Washington State College; Paul R. Hanna, Stanford University; Albert H. Marckwardt, Princeton University; Harold E. Mitzel, Pennsylvania State University; William McColly, University of Wisconsin; Neil Postman, New York University; Marjorie Smiley, Hunter College of the City of New York; Ruth G. Strickland, Indiana University; and Frank J. Zidonis, Ohio State University.

Some members of the group registered a demurrer to some of the grandiose connotations of the term *research design.* They argued that more needs to be known about the nature of language and the learning process before studies may be designed dealing with *teaching language.* The majority of the group agreed that:

1. Research design does not necessarily imply a large-scale, multivariate investigation; that *informal* and *micro-* studies are appropriate and desirable prerequisites to more elaborate studies. Such microstudies and informal investigations, however, must be carefully done and communicable in nature if they are to be useful.
2. The basic principles of research design apply equally to small-scale and large-scale studies. Principles stressed included (a) measurement of outcome independent of the teacher's or experimenter's conclusion; (b) control of extraneous variables, such as students' previous educational backgrounds, intelligence, and cultural experiences; and (c) isolation of the particular variable whose effect is being investigated.
3. The results of informal studies, such as research conducted by the classroom teacher, should be so formulated as to lead to hypotheses which can be tested in formal research projects.

Considerable attention was given to the course, curriculum, or program as the major variable in studies. From this discussion the following conclusions appear to be pertinent:

1. A curriculum is actually a composite of many elements. Although it has often been treated as a variable in experiments (e.g., experimental and control group comparisons), it is not a unitary one. Thus it is difficult to learn what curriculum elements are responsible for a particular observed change, or whether elements might be cancelling each other out.
2. The "real" curriculum is a function of the behavior of the teachers and the students; this behavior is often unknown to the curriculum evaluators. Thus, teachers' diaries, interviews, and classroom observations may need to be used more frequently to "keep track" of the curriculum.

A brief amount of time was given to considering approaches to research design which would take into account the above conclusions and and principles. One approach that was suggested involved sequential measurement of selected aspects of child performance at intervals throughout a period of time. The results of such measurement might then be plotted graphically against time. Some of the problems that might be approached in this way are (a) study of the growth of individuals without intentional manipulation of the curriculum; (b) comparison of such growth with that resulting when an experimental program is introduced; and (c) study of changes in the shape of the learning curve when a new curriculum element is introduced.

Another proposal was based on the assumption that, if research could determine that there are specific kinds of sentences which pupils have trouble identifying and using at different age levels, this information would have great import for instruction. Several means were suggested for determining the knowledge and ability that children have. These included (a) the recording and analyzing of children's speech, using a generative grammar as the basis for such analysis; (b) testing children's ability to decide whether certain words and phrases could be put together and make sense or not; and (c) presenting children with a list of sentences, graded from the least to the most difficult and interspersed with nonsentences, to test their ability to recognize sentences.

A fourth suggestion for the measurement of language learning was by galvanic skin response. It was suggested that the galvanic skin response is greater upon presentation of semantically related stimuli than of those phonologically related (e.g., "doctor-physician" than "doctor-proctor"), and that it is likely that normal children will show a greater galvanic skin response to an internalized (learned) structural relationship or concept than to those not known or not structurally related.

The group also considered a proposal dealing with the time in a child's school life when various concepts about language could be economically grasped. This proposal calls for the devising of exercises suitable to the various grade levels of the children studied. These exercises would be of progressive complexity and would involve such concepts as base, superfix, word; free form, bound form; prefix, suffix; derivational and inflectional suffixes; and perhaps some allomorphs and some morphophonemic changes.

Other proposals advanced but not developed included:

1. Studying the teaching of an artificial language by means of deductive and inductive approaches.
2. Analyzing the effects of different amounts and kinds of reading upon language patterns.
3. Comparing the use of two or more descriptive systems for teaching the nature of language.
4. Applying the principles of second-language learning to the learning of standard English.

The group concluded its study by stressing the importance of cooperative planning and research by the specialist in design, the specialist in language, and the specialist in teaching.

Research in Teaching Literature[2]

After considering a number of possible starting points for their deliberations, the members of the group divided into three smaller groups each to define a problem in the teaching of literature and to outline a design for research on the problem.

Subgroup A elected to explore which aspects (in a Shakespearean play) can best be taught by the lecture method—character, plot, theme, or imagery. The proposed design was a rotation plan with three teachers each teaching three Shakespearean plays to student groups equated on (1) language ability (English test, reading test); (2) verbal intelligence; and, possibly, (3) socioeconomic status. The number of lessons would be kept identical for all three groups and methods.

Each group would be taught by three methods: A—lecture only; B—discussion only; and C—combination, perhaps half lecture and half discussion. Thus, the first play would be taught to Group 1 by Method A, to Group 2 by Method B, and to Group 3 by Method C. Tape recordings or observations would be utilized to confirm that the specified methods were actually used in each group.

Following the teaching of a specified number of lessons, a test composed of objective and essay items to evaluate character, plot, theme, and imagery would be constructed and administered. The test would contain many short items rather than a few long essay questions, mixing items relating to various aspects (plot, theme, etc.).

The second and third play would each be taught to these same student groups, but with a rotation of the methods, as shown below. At the end of each series of lessons, a test comparable to that given for the first play would be administered.

	Group 1	Group 2	Group 3
Play 1	Method A	Method B	Method C
Play 2	Method B	Method C	Method A
Play 3	Method C	Method A	Method B

Each teacher would use each method, each play would be taught by each method, and each group would use each method and play.

---

[2] **Group Roster: Leader,** William H. Evans, University of Illinois; **Recorder,** Patrick J. Grofl, San Diego State College; Edwin A. Peel, University of Birmingham; G. L. Anderson, Modern Language Association of America; Jerome W. Archer, Arizona State University; John A. Brownell, Claremont Graduate School; Bruce Harkness, Southern Illinois University; Helen Huus, University of Pennsylvania; John E. Jordan, University of California; Elliott Landau, University of Utah; Lewis Leary, Columbia University; Henry C. Meckel, San Jose State College; Norine Odland, University of Minnesota; Robert W. Rogers, University of Illinois; Louise M. Rosenblatt, New York University; Kester Svendsen, University of Oregon; Erwin R. Steinberg, Carnegie Institute of Technology; James R. Wilson, San Francisco State College.

The final step would be to have a fourth play read independently and evaluated by a test, comparable to the others, over character, plot, theme, and imagery.

An adaptation of this design would be to rotate teachers, that is, to have each teacher use only one method but to proceed from group to group. Other variations were suggested: lecture with and without visual aids, teams of teachers rather than individual teachers, groups of different sizes, lectures via television versus "live" teaching, emphasis on other aspects of literature than the four mentioned, e.g., symbolic structure, style, relationships, setting, tenor, interests, and attitudes. Several schools and colleges might use the same design and pool their data. Despite the simplicity and adaptability of the design, it would have limitations: (1) the groups could be difficult to equate; (2) the teachers may have difficulty maintaining "pure" lecture and "pure" discussion methods; (3) one play may best be taught by one method; (4) the order of the presentation of the methods may influence their effectiveness; understanding and preference of individual teachers undoubtedly would vary for the different plays.

Subgroup B proposed a study to determine the relationships between, on the one hand, the stated aims of a teacher for teaching a piece of literature and, on the other, the kinds of questions he asks and the oral and written assignments he makes. The study would (1) gather from teachers a large, representative sample of the kinds of questions they ask and oral and written assignments they make for certain selected literary works; (2) obtain the objectives or aims teachers say they have in teaching these literary works; and (3) compare these stated aims with the questions and assignments. The variables of the study would be explored through classroom observations, questionnaires the teachers respond to, interviews with the teachers, and samples of examination questions used by the teachers. A selected group of literary experts would be assigned to determine the disparity or agreement of the stated aims, the questions asked, and the assignments made.

The degree of agreement of the aims and questions and assignments would be studied in relation to the nature and extent of the teacher's education, his sex and age, the length of time he had taught, the nature and location of his school, his teaching load and extracurricular responsibilities, his willingness to have the data on his teaching practices gathered, and the specificity with which he stated his aims for teaching literature.

Subgroup C proposed to compare depth versus breadth of reading on the effectiveness of studying the short story. The study would have one teacher spend two weeks with a class on a single short story, a second teacher spend two weeks with a class on ten short stories, and a third teacher spend two weeks with a class on five short stories. Then, an evaluation would be made of how well the students in each of the three classes deal with a new short story.

The group considered several means of measuring the effectiveness of experimental variables in the teaching of literature: a short poem could be rewritten to flatten out all the images or turn them into clichés. As much as possible, metrics and other aspects would be kept the same. Both poems would be presented to the students, who would be asked to select the better poem and to define the choice.

Students would be confronted with a short poem (perhaps a sonnet) in which one of the lines is underscored, and a series of interpretations is offered for the underscored line. The students would rank the alternatives in the order of the best interpretation and defend in writing their choice of the "best" interpretation. Each student's order of preference would be compared with an order determined by a panel of experts, and the written defenses would be evaluated.

An irony scale could be established—a series of ironic passages arranged in increasing order of sophistication. The scale would be used to establish norms for various grade levels. Once the norms are established, a teacher may determine how well a class or individuals in a class perform according to the norms.

Research in Teaching Composition[3]

The meetings of Group III might best be described as exploratory, resulting primarily in an exchange of views and information. The group decided to consider the products and processes of composition as well as research in the teaching of composition. A reasonable approach, it was felt, was one which would separate problem areas into the methodological, substantive, and logistical. Consensus supported the development of a taxonomy in composition teaching.

The group saw merit in supporting a series of related, small experimental or pilot studies, limiting the numbers of variables in each situation, and to replicating under local conditions studies already reported. Attitude studies were discussed at some length.

Some group members thought that the rewards accruing from such studies are too limited to justify the necessary time and effort; others conjectured that the accumulation of data concerning attitudes toward writing might eventually prove of far more consequence than one is now able to estimate.

---

[3] **Group Roster: Leader,** Richard Braddock, University of Iowa; **Recorder,** Robert L. Wright, Michigan State University; Alvina Treut Burrows, New York University; Paul B. Diederich, Educational Testing Service; Edna Furness, Nebraska State College, Kearney; N. L. Gage, Stanford University; Robert M. Gorrell, University of Nevada; Arno Jewett, U. S. Office of Education; Silvy Kraus, University of Oregon; Edwin H. Sauer, Chicago Teachers College South; Ingrid Strom, Indiana University; Anthony Tovatt, Ball State University; Hazel Browne Williams, University of Missouri, Kansas City.

The group also expressed interest in hypotheses which research in reading may suggest for research in written composition. Several members stated that present composition research has placed too much emphasis on the effects of student writing upon teachers and raters rather than upon the understanding, emotion, and behavior of other types of readers. Other members called for more research focusing on descriptions of specific rhetorical considerations and fewer general evaluations of entire compositions.

Among other topics discussed were observational techniques, case study analyses, factors of maturation as related to writing, kinds of control groups, effects of paper revision, the use of equivalent theme topics, and types of validity checks. The group regarded longitudinal studies as particularly worthwhile.

The group recommended that the Buxton study be replicated.[4] Similarly, interest in, but no unanimous support for, some of the following proposals was evident:

1. Encouraging experimentation in composition by providing financial and professional support.
2. Creating an information center or agency through which researchers could learn of others working in similar areas or engaged in related investigations, communicating with specialists of various kinds, finding out about studies underway, asking for advice about research design, and so on.
3. Publishing regular reports not limited as to scholastic level on research in written composition.
4. Appointing of an NCTE official whose function would be to facilitate the three points mentioned above.

No model for research design grew out of the meetings of Group III. Such an expectation may not be completely realistic without greater commonality of interest in a specific project and more thorough premeeting preparation of or by participants.

---

[4] Earl W. Buxton, "An Experiment to Test the Effects of Writing Frequency and Guided Practice upon Students' Skill in Written Expression" (Unpublished Ph.D. dissertation, Stanford University, 1958). Summarized in detail in Richard Braddock, Richard Lloyd-Jones, and Lowell Schoer, **Research in Written Composition** (Champaign, Illinois: National Council of Teachers of English, 1963), pp. 58-70.

Research in Teaching Reading[5]

The study group on the Teaching of Reading considered the problems of research in reading in relation to three topics: (1) the forthcoming USOE sponsored studies of reading methods in grade one; (2) limitations of research design, especially in relation to methods studies; and (3) the relationship and contributions of linguistics to early reading instruction.

During much of the first meeting of the committee, attention was focused on the proposed studies of beginning reading. Two basic issues were considered:

1. The nature and design of specific proposals.
2. The need for a coordinating center which would serve the various researchers by helping them:
    a. to determine the adequacy of research design before the studies begin. (Mr. Harris suggested that the design include the tables used in presenting this data to ensure that the studies would answer the questions raised.)
    b. to develop guidelines for test selection, statistical measures to be applied in the analysis of data, provision for control groups, methods by which the experiment could be considered periodically to determine important changes occurring during the study. (The consultant on research design raised important questions related to the study of continuing effects resulting from varied treatment and the need to know exactly what goes on during the tenure of the experiment in both control and experimental groups. During the one of the general sessions, Mr. Gage made a similar plea concerning the need to know what was taking place during experiments, for example, between pretesting and post-testing.
    c. to examine the possible implications for other cooperative studies which might grow out of the first grade studies.

During the discussion of the first grade studies, members of the group voiced considerable trepidation about the short duration (one year) of the proposed investigations. Several persons suggested that most first grade instruction in reading could not be adequately assessed unless follow-up studies with interim observations of pupils and teachers were carried out for as long as three and four years after the initial year of experimentation.

---

[5] **Group Roster: Leader,** Theodore Clymer, University of Minnesota; **Recorder,** William D. Sheldon, Syracuse University; Chester W. Harris, University of Wisconsin; David Reed, University of California; Dolores Durkin, Teachers College, Columbia University; Donald Durreli, Boston University; Leo Fay, Indiana University; Henry R. Fea, University of Washington; Oscar M. Haugh, University of Kansas; Jack A. Holmes, University of California; Thomas Horn, University of Texas; William Iverson, Stanford University; Helen M. Robinson, University of Chicago; Robert E. Shafer, Teachers College, Columbia University.

Mr. Harris made a plea for clarifying objectives in each of the studies. He also stated the need for replication of studies in order to give more substance to generalizations drawn from studies. It is this need for replication that caused him to suggest the following points:

1. The research worker needs to spell out the daily or weekly happenings or the actual interaction between pupils and teachers in methods studies. He needs to understand those measurable aspects of a study which will enable him to provide in advance for the proper tools of measurement and which will allow recording of those significant reactions which are so often overlooked when research design is unsophisticated. For example, the use of cameras, concealed or exposed, helps portray specific behavior in a reading class, especially during direct instruction on the one hand or during unsupervised activities on the other. Tapes of various oral reading activities or post-reading discussions also provide an excellent body of analyzable data, significant to the researcher.

   Further observers might play an important part in evaluating ongoing research. Teams of objective, well-trained observers would be invaluable assets in aiding the research worker to describe what had actually taken place during his study.

2. The research worker needs to be close enough to the experiment to ascertain what is happening and under what conditions it happened.

3. The researcher also needs to be able to conceptualize as to what was being done.

4. The investigator must be continuously aware of the treatment variables.

5. The researcher needs to be able to answer questions related to the different results which might have been obtained through different treatments. (This is particularly pertinent to the first grade study, where the varying treatments need to be clearly defined and carefully applied.)

The consultant on design also cautioned the group about being fully aware of the relationship between dependent and independent variables. A helpful discussion on the stratifying of variables ensued. The group elicited from Mr. Harris not only a series of basic design considerations but also bibliographical references which might aid the designers of studies.

Because the group had been discussing a specific area of concern in its first session, the chairman during the opening moments of the second day of meeting reminded the committee of the purposes of the conference. He suggested that the group divide into three working groups which would attempt to identify designs useful for research in the teaching of reading and develop proposals illustrative of the designs. However, the group, wishing to continue discussion with Mr. Harris on the limitation of current research designs and how new designs might be developed, decided not to break into groups.

Later Mr. Reed was asked to discuss the contribution linguists might make to reading instruction. After suggesting what contributions he believed linguists might make, Mr. Reed then pointed out the linguists' limited understanding of classroom instruction in reading at the primary grade levels.

In the third session, Mr. Reed commented on the paper which reported on the Interdisciplinary Seminar in Linguistics and Reading held in Seattle during the summer of 1963. This project had as one of its objectives the creation of materials based on a linguistic point of view which would be used in early reading instruction. Instead, it achieved a statement which helps clarify the current differences in the points of view of some linguists and reading specialists.

The group was disappointed because it was unable to devise a paradigm which might serve the research worker in reading. However, each member of the committee stated that the conference had enlightened him about the need for improved research design and each was thankful for the many suggestions made by the research experts.

Research in Communication and Verbal Learning[6]

The study group in communication and verbal learning was faced with a two-fold problem: considerable diversity in the background and specific interests of its members, and corresponding variances in individual conceptions of research which would appropriately differentiate the work of this group from that of the other study groups.

The wide divergence of the group became apparent in the initial study session in which it was suggested that the group move toward the formulation of a model research design through an examination of the sorts of hypotheses and designs actually used in basic research in communication and verbal learning. After considering one specific example of a published experimental study of oral communication,[7] group members raised many questions in an effort to clarify their task. Accordingly, each member agreed to submit questions which might be significant and appropriate for developing one or more research designs by the entire group.

---

[6] **Group Roster: Leader,** John B. Carroll, Harvard University; **Consultant,** Leo J. Postman,
University of California; **Recorder,** Theodore L. Harris, University of Wisconsin; John DeBoer,
University of Illinois; William E. Hoth, Wayne State University; Marie Hughes, University
of Arizona; Arnold Lazarus, Purdue University; Walter H. MacGinitie, Teachers College, Co-
lumbia University; Mary Regan, University of California; Hilda Taba, San Francisco State
College; Joseph Thomson, University of Illinois; Francis Shoemaker, Teachers College, Colum-
bia University; Bernice Wolfson, University of Wisconsin-Milwaukee.

The questions submitted appeared to fall into two categories: (1) those having to do with the effects of *affect* and *feedback* upon the learning and/or teaching of verbal materials, and (2) those having to do with the effects of *context* upon the learning of new words and upon the difficulty of decoding a communication. Because suggested problems range from the highly specific to the very general, it was clear that no single pattern of research design would be applicable to all proposed.

After discussing the difficulties in controlling *affect* in an experimental investigation, the group decided that a study of *context* effects upon the learning of new words might be more feasible.

The group next discussed a study of the effect of linguistic structure upon the recognition of a new word presented via the medium of an artificial language. Attention was given to the definition of the problem, to relevant related research, and to questions of design and measurement.

At this point, the group leader generalized the problem of learning new words to that of ways of building vocabulary. The group responded positively to a suggested focus upon ways of building children's understanding of the various meanings of complex words.

More specifically, the topic proposed was an examination of the role of inductive and deductive procedures in vocabulary building. A wide range of variables which might be examined in such study was discussed and several tentative research designs suggested. The group soon recognized, however, that the problem was too complex for a highly detailed research design to be developed in the time remaining. It was therefore decided that the proposal should identify the general problem area, include a simplified illustrative research design based upon selected aspects of the group's discussions, and indicate a number of additional variables which might be incorporated within the general framework of the proposed design.

PROPOSED RESEARCH DESIGN

A STUDY OF INDUCTIVE AND DEDUCTIVE PROCEDURES IN VOCABULARY BUILDING

Introduction. Vocabulary building is a prime responsibility of the teacher of English. Yet he has little objective evidence about how vocabulary meaning is most effectively developed, nor does he know precisely how successful children are in recognizing the essential attributes of rather complex verbal concepts. This study proposes a way of exploring the effects of two types of controlled presentation of illustrative verbal contexts, inductive and deductive, upon children's understanding of the meanings of words with several defining attributes.

---

[7] Howard Maclay and Stanley Newman, "Two Variables Affecting the Message in Communication," in Dorothy K. Wilner (ed.), **Decisions, Values, and Groups** (New York: Pergamon Press, 1960).

Relatively recent work by Bruner and others in concept attainment tasks suggests an experimental approach which may usefully introduce some precision and objectivity to such a study. In effect, this approach involves presenting a student with an array of positive and negative examples of a concept such that each example either contains or does not contain defining attributes of the concept to be attained. The experimental task, for example, might be arranged so that the student would be asked to identify appropriate examples, one at a time, of a given concept from the array of examples before him. After each choice, the student would be told whether his choice was or was not correct, i.e., whether it represented an appropriate attribute of the concept. The pattern of choices and the speed with which a concept is attained, i.e., its essential attributes identified, could then be objectively analyzed.

In most of the research conducted under such experimental conditions, geometric figures (presented on a display board) have been used as stimuli to represent various combinations of such attributes as size, shape, color, etc. Verbal materials, e.g., a series of contexts describing certain attributes of a verbal concept such as a word, might be similarly displayed before the student or presented successively on cards. His success in understanding the contexts (attributes) relevant to the verbal concept might then be examined under two conditions of presentation: a *deductive* task of applying the definition of a word in the identification of correct verbal contexts or examples of its meanings, and an *inductive* task of identifying the word or verbal concept through the examination of appropriate verbal contexts or examples.

The general approach suggested here might be used to design not just one but a series of studies. Therefore, an illustrative model will be first described, and then possible extensions and refinements of this general model suggested.

Problem. To investigate the role of inductive and deductive procedures in vocabulary building at the sixth grade level.

Population. One hundred and twenty-eight sixth grade children of average verbal ability (IQ 90-110), at or above grade level in reading ability.

General Methodology. Present subjects with a random array of illustrative verbal contexts under two task conditions, inductive and deductive. A group of related contexts would represent correct or incorrect examples of the defining attributes of a single-word concept. (Several such arrays might be developed, one for each concept selected, and placed upon cards. Concepts of some complexity, i.e., having a number of defining attributes, would be used. Each illustrative context developed would be of the same length.)

Specific Task. (a) *Preparation.* Construct twenty illustrative verbal contexts of twelve words each for each of two words: (1) *climate* and (2) *compromise,* plus an additional negative and a positive context for each concept to be used for training purposes in the inductive task condition.[8] Half of the contexts should be positive (including one or more defining attributes) and half negative (including incorrect and/or no defining attributes). The cards should be randomly shuffled.

(b) *Assignment of Subjects to Tasks.* Assign subjects randomly but equally by sex to one of four task conditions:

Condition	Subjects		Concept 1	Concept 2
	M	F		
1	16	16	Deductive	Deductive
2	16	16	Deductive	Inductive
3	16	16	Inductive	Deductive
4	16	16	Inductive	Inductive

Each subject would attempt both concepts under the one task condition to which he is assigned.

(c) *Procedure: Deductive Task.*

(1) Present definition of concept on card. Ask subject to read definition aloud to verify recognition of words used.

(2) Present twenty contexts on cards in random order. Ask subject to say "yes" if the card represents a correct meaning of the definition, or "no" if a meaning is incorrect.

(3) Confirm each response "yes" or "no."

(4) Score: time taken to attain ten correct identifications in order. (If necessary, reshuffle cards not correctly identified and present again.)

(d) *Procedure: Inductive Task.*

(1) Tell subject he is to try to identify the concept from examples of its meaning, some right and some wrong, presented on cards. Ask him to guess the correct concept after each card is presented.

(2) Present positive training example: confirm.

(3) Present negative training example: confirm.

(4) Present twenty cards in random order: confirm.

---

[8] Other concepts suggested by the group for possible use include such concepts as asset, maturity, comedy, mammal, frontier, neighbor, disgrace, and cooperation.

(5) Score: time taken to achieve ten correct identifications *after and including* the first correct identification of the concept. (If necessary, reshuffle cards not correctly identified and present again.)

(e) *Analysis.* Analysis of variance to determine treatment, order and sex effects, and the interaction among these variables.

PROPOSED EXTENSIONS AND MODIFICATIONS OF THE DESIGN

This research design may be modified and extended in several ways to study the effect of other variables or combinations of variables upon the major inductive and deductive process variables incorporated in the model design. Among the variables which might be introduced with appropriate adjustments in the design are the following:

Selection Variables

1. Different grade levels, separately or in combination
2. Different levels of academic achievement
3. Different levels of intellectual ability
4. Different socioeconomic levels
5. Different combinations of the above selection variables

Treatment Variables

1. Different types of pretraining and/or instruction to subjects
2. Different length of tasks involving different degrees of memory strain

Stimulus Variables

1. Use of display board for presentation of all contexts simultaneously
2. Various degrees of complexity and/or abstractness of concepts and illustrative contexts
3. Planned, rather than random, ordering of positive and negative examples
4. Use of concepts presumably different in degree of affective connotation (*resentment* versus *lake*)
5. Use of partially correct as well as correct and incorrect examples

Further exploration of the effect of such variables upon two primary ways of building a vocabulary, through inductive or deductive procedures, should add greatly to our understanding of how to help children in turn expand their understanding of the meanings of complex word concepts.

PROJECT ENGLISH — A LOOK AHEAD

Francis A. J. Ianni

I suspect that I was asked to make this concluding talk at the research conference on English for several reasons. One obvious one is that I represent the financial host. A second reason, however, is that my colleagues in English tell me that I am supposed to have an approach to such speaking assignments, an approach which has variously been characterized as a lightness of touch, a lightness of heart, and even a lightness of head. The events of the past few days, however, have done away with any feeling of frivolity or even good humor on my part. The senseless and tragic act which took from us both a beloved President and a leading exponent of American education has left me devoid of any such feelings. I must ask you to excuse then the fact that I felt compelled to put aside the somewhat lighthearted speech I had prepared and to allow me to substitute these few remarks.

The Future of Project English

Any attempt to look at the future of English recalls to me the ancient Chinese proverb which asks if there are more fools made by looking into the future than by looking into the past. My own answer is that there are really only two types of complete fools: the fool who looks at the future with no reference to the past and the fool who dwells on the past and has no time to look into the future. It is impossible to discuss the future of Project English without some look at the life history of the approach to educational research which has led to Project English. Let me then take some few moments to describe to you how Project English came about.

Project English was born in the thrilling early days of the new frontier. It was the expression of a feeling—so well characterized by the restless spirit of the then new President—which said to the educational community, "Let's get moving." It was based upon a concept that research should be more than counting, classification, and correlating, that teaching should be more than precepts, platitudes, and pedantry. Finally, it was an expression of the firm belief that federal support of education is not synonymous with federal control. President Kennedy expressed this last view in one of his early messages:

> The control and operation of education in America must remain the responsibility of state and local governments and private institutions . . . But the Congress has long recognized the responsibility of the nation as a whole—and additional resources, meaningful encouragement, and vigorous leadership must be added to the total effort by the federal government if we are to meet the task before us. For education in this country is the right, the necessity, and the responsibility of all. Its advancement is essential to national objectives and dependent on the greater financial resources available at the national level.

The response was to find some means of reducing the lag between research and practice. Early in 1961 a new administration asked for a new approach that would (1) develop new materials for the schools; (2) translate research into action; (3) inject innovation and vigor into the nation's schools.

The President's budget charge delivered to us at that point voiced his concern with the status of education and the special needs of the scholarly pursuit:

> Our progress as a nation can be no swifter than our progress in education. Our requirements for world leadership, our hopes for economic growth, and the demand of citizenship itself in areas such as this all require the maximum development of every young American's capacity. Our twin goals must be: a new standard of excellence in all areas of education and the availability of such excellence to all who are willing and able to pursue it.

Our response to this challenge was to plan some method of mobilizing the resources of the Office of Education and the educational community into a new attack on the pressing educational problems of our day. In addition, we hoped to develop new approaches to educational research and development which would bring together the scholar, the researcher, the teacher, the administrator, the student, and all who are involved in the vital process of education. The area of English was chosen as the model for future development. From this period of almost chaotic excitement and challenge was born Project English.

Our first beginnings were—like all experimental programs—a matter of trial and error. Where could we turn for the kinds of information that we needed? The Office of Education had traditionally turned to the schools of education. The National Science Foundation, on the other hand, had turned much more frequently to the university scholar and had sought to bring change into education by involving the physicist, the biologist, and the chemist from the university in planning educational change in the high schools and elementary schools of the nation. We sought to combine both approaches, and, in doing so, we suffered from the attempt to bring together two different streams. In our naivete we expected to be blessed by both the educator and the scholar, while in reality we were roundly condemned by both. We took information from all who would give it to us. And at that time I could not help but remember Swift's satiric analogy in *The Battle of the Books.* As you are well aware, this was the 18th century argument over the comparative virtues of the ancients versus the moderns. The spider, who represents the moderns, brags that he secretes everything from his own nature, and therefore, everything comes out of imagination. The bee on the other hand says, "I take from one flower, then another, then another, and then I create something from my own nature, something of sweetness and light." And so we took from everyone suggestions for Project English, trying to synthesize them into our own new approach.

Project English, however, was not the simple result of this brainstorming session in the Office of Education. Rather, it had roots in both the immediate and far distant past. Historically, it was certainly a part of the long-standing tradition of interest the federal government has shown in the affairs of education. To recognize the extensiveness of the federal concern, one need only recall the legislation which established the Smithsonian Institution in 1846, the land grant colleges in 1862, the Office itself in 1867, and more recently the National Science Foundation in 1950, the Cooperative Research Program in 1954, and the National Defense Education Act of 1958.

A closer look at the history of the Office of Education, however, suggests that Project English was not a natural outgrowth of the emerging role of the Office. The enabling act of 1867 describes the purposes of the Office as " . . . collecting such statistics and facts as shall show the condition and progress of education. . . . and of diffusing such information as shall aid the people of the United States in establishing and maintaining efficient school systems, and otherwise promote the cause of education."

This enabling legislation had to be supplemented over the years by various pieces of legislation that have authorized the Office to (1) conduct studies on various aspects of education; (2) provide financial support for extramural research and development; (3) coordinate the teacher exchange program, training programs for foreign educators, and other programs with foreign countries; (4) provide grants and loans to state departments of education and educational institutions for the improvement of educational practice; (5) disseminate statistical information, research findings, and descriptions of promising new instructional practices and programs.

Much had to be added before Project English could become effective. While it would be less than honest, and even presumptuous, to suggest that Project English represented the same concern to the new frontier that the Peace Corps, civil rights, the Alliance for Progress, and other action-oriented programs held, it was nevertheless an important enterprise in the Administration program. The heart of Mr. Kennedy was in the type of approach represented by Project English. In an early budget message, commenting on areas for emphasis in future budgets, he said:

> I regret very much the failure of the Congress to enact legislation for general aid to education. I am heartened by the success of our expenditures in science and research. I feel that the new endeavors in educational research will be of the utmost importance to the future growth and strength of our nation.

With new impetus and with the existing authority of the Office of Education, we started with a total budget of four million, of which five hundred thousand was for English. In the two years that Project Eng- has been in operation, our budget has grown to twelve million, with two and a half million for English. We began Project English with three types of programs: basic and applied research projects, curriculum study centers, and conferences and seminars.

To these three programs which were available in 1962, we have added in 1963 two new programs—small contracts and research and development centers—and an expanded demonstration program.

Small Contracts

The small contract program will support small-scale research and development— projects requiring $7,500 or less in direct federal support. This program has been designed to encourage doctoral students to do more significant research than they could do without assistance and to enable experienced researchers to probe a problem, to look into the potential value or the feasibility of an extensive research project, without presenting a fully developed research design. Recipients range from students to Nobel Prize winners.

Proposals under this program have the advantage of a special review process which takes less time than the regular procedure. There are no deadline dates; review begins on receipt of the proposal, and funds can be made available for an approved project about two months after the proposal is received in the Office. Any students wishing to submit a proposal under this program must be sponsored by a staff member of a college, university, or state department of education who will assume responsibility for seeing that the research is completed according to the terms of the contract.

Research and Demonstration Centers

In an attempt to make a significant contribution toward the understanding and improvement of educational practice in particular problem areas, research and development centers will be established during this fiscal year to concentrate extensive human and financial resources. The activities at each center will include basic and applied research studies; developmental activities designed to translate research findings into educational materials and procedures; and demonstrations of, and dissemination of information on, the newly developed and tested materials and procedures. In addition, the professional staff of the center will provide leadership in the chosen problem area. Although these centers were not originally conceived as subject matter oriented, this kind of orientation was not excluded from consideration, and hence it would be appropriate to consider support for a research and development center in English.

Demonstration Programs

A greatly expanded demonstration program rounds out the types of support now available. The demonstration is a natural follow-up to the basic and applied research studies and the curriculum improvement projects. Its purpose is to display to the profession the new materials, methods, techniques, and programs that have been proven effective through research. A demonstration may make use of the live classroom setting, educational television, radio, and other communication media. Our hope is that the demonstration program will reduce the lag which exists between what we know through research and the what and how of teaching in the classroom.

What of the results of Project English? In one sense, it is premature to talk about results with so young a project. After all, there have been no revolutionary changes in instructional programs in the schools; no new materials have emerged; no new content has been identified. But in another sense, a number of results have already begun to emerge. Let me briefly enumerate them.

1. The educators and academicians have sat down together and have found that in most respects their differences are no greater than differences within their own groups. Each group has begun to develop a respect for the thinking of the other, and each now appears to be willing to examine common problems rather than faults. We are most proud of the fact that those of us in the Office are parties to this partnership.

2. Larger sums of money are being made available for the study of the English program. While no one would be so naive as to equate money with progress, the latter is almost impossible without the former.

3. The expansion in funds has been accompanied by an expansion of the types of support programs available.

4. The interest of the profession including the academic disciplines has been aroused not only in English but in other disciplines as well.

5. A total of ten curriculum centers, more than seventy research projects, four demonstration centers, and nine developmental activities have received support.

Project English represents a progress which is much more meaningful to us than the gross in a number of projects and in the amount of money given to us by the Congress. But there are other measures of progress which are even more meaningful to us. Let me repeat that I think the communication established between the English scholar and the English educator is one of the great successes of Project English, for it has now become respectable for the English scholar to turn his attention to education.

Finally, Project English has produced a vision of action for the future.

One year ago in Miami I listed fifteen new needs for Project English. Since then, we have realized five of these fifteen needs. But ten still remain unfulfilled, which we hope to meet in the coming year. Let me list these ten for you once again.

1. The establishment of a number of model demonstration schools where new materials can be demonstrated to practitioners.

2. The establishment of a number of endowed chairs at smaller colleges and universities so that specialists in research can work with the staff members of the smaller schools in developing adequate research opportunities.

3. An international exchange of researchers and scholars working in the field of education.

4. A large number of institutes and support for group study as a means of improving the teaching of English in both preservice and inservice programs.

5. The development of a system of institutional support so that colleges and universities can build their own research potential.

6. Expansion of our present research and development center program to at least fifty centers so that each state will have the opportunity of using new approaches to education.

7. Grants for the purpose of purchasing necessary research and development equipment for universities.

8. Pre- and post-doctoral fellowships to support graduate students, and advanced scholars in the field of English.

9. A program of career grants so that established researchers can devote the remaining years of their lives to research without needing to apply regularly for additional funds.

10. A national institute program in the area of education which would parallel the existing National Institutes of Health.

We still have a great distance to go if we are to bring to realization our hopes and dreams for English. We have made only a few halting steps in that direction, but as President Kennedy often pointed out, even a trip of a thousand miles must begin with the first step.

Perhaps the greatest need is a growth and continuation of the existing cooperation between the field of English and the Office of Education. This has not always been a comfortable relationship, for in many ways the scholar resents the association with the bureaucrat and the politician. I can think of no better illustration of this problem than that supplied by President Kennedy on February 26, 1961, for the television program on Robert Frost:

> There is a story that some years ago an interested mother wrote to a principal of a school, "Don't teach my boy poetry. He is going to run for Congress."
>
> I have never taken the view that the world of politics and the world of poetry are so far apart. I think politicians and poets share at least one thing, and that is that their greatness depends upon the courage with which they face the challenges of life. There are many kinds of courage—bravery under fire, courage to risk reputation, friendship, and career for convictions which are deeply held. Perhaps the rarest courage of all—the skill to pursue it is given to very few men—is the courage to wage a silent battle to illuminate the nature of man and the world in which he lives.

EVALUATION AND IMPLICATIONS

Margaret J. Early, David H. Russell, and James R. Squire

To put the question directly: Was the San Francisco Conference worth the expenditure of time and money to the seventy-five participants and to the sponsoring agents? What, if anything, did it contribute to the state of research in the teaching of English? Although direct answers are not so easily formulated as direct questions, this final chapter examines outcomes and speculates upon next steps. To a degree, this is a subjective evaluation by the writers, a statement of their impressions; but it is also, we believe, a reflection of the ideas expressed by the recorders and consultants at the final session. In particular, we are indebted to N. L. Gage for expressing informally at that session some of the ideas included below under "Future Directions."

What Purposes Were Served?

Perhaps the best basis for evaluation of the conference lies in these proceedings, and each reader, like every participant, will make his own direct answers to the questions raised above. In doing so, the reader will ask the classic critical questions: What were the purposes? How well were they achieved? Were they worth pursuing? It would be presumptuous for the writers to try to answer these questions *for* the reader; nevertheless, we propose to use them as a handy guide to the evaluation, not of the proceedings alone, but of the conference itself.

Some of the purposes of the conference were stated by Mr. Russell in his opening remarks. The purposes as they emerged more formally in the various stages of planning were these:

- –To further explore ideas for research in English that emerged from previous conferences, going beyond statements of needed research to a concern with problems of research design.
- –To examine methods of research in literary scholarship and linguistic science, in methods of teaching, in verbal learning, and in other aspects of educational psychology; and to consider the implications of these for research in the teaching of English.
- –To consult with experts in research design and statistics on specific problems of research in five areas related to teaching English.
- –To test, by the example of the conference itself, in general ses-

sions and in small working groups, the hypothesis that experts in literature and language and in methods of teaching and the psychology of learning can contribute to each other's thinking and, more than that, can work together effectively to improve the state of knowledge about teaching English.

To accomplish these purposes, the conference planners proposed two measures. The first was to invite twelve representatives of disciplines named above to prepare papers for presentation at general sessions and for inclusion in the conference proceedings. Thus, Mr. Steinberg and Miss Rosenblatt summarized and evaluated the outcomes of two previous conferences and suggested directions for the present one. Mr. Gage described paradigms for research, designated as substantive, methodological, and logistical, and related them to the field of English. Mr. Ianni reviewed research and curriculum projects supported by the United States Office of Education and forecast plans for the immediate future. Three consultants in research design—Messrs. Mitzel, Peel, and Harris—described techniques and commented on developments in three aspects of research methodology: classroom observation, analysis of preferences, and factor analysis. Two scholars in the study of literature and language, Mr. Leary and Mr. Marckwardt, commented upon methods of research characteristic of their special fields. Mr. Meckel, representing English education, surveyed the need for studies of sequences of learning and teaching particularly in the secondary school; Mr. Durrell focused upon studies of reading and language arts in the elementary school; and Mr. Diederich examined the complex problems of evaluation in a specific field—written composition.

In addition to these papers, general sessions were planned in which the four research consultants reacted to ideas for substantive and methodological research introduced by Meckel, Durrell, Marckwardt, and Leary. At the final session, the study group recorders, each a member of the Committee on Research of the National Council of Teachers of English, expressed their views of the conference as a whole and the deliberations of their respective study groups. Again the research consultants were called upon to evaluate the conference and to suggest next steps.

As a second means of pursuing the objectives, the conference planners set up five study groups, assigning to each one consultants in research design, scholars in literature and language, and specialists in teaching. To give focus to the discussions and indeed to demonstrate the process of cooperative planning, it was suggested that the leaders present examples of pertinent investigations for critical examination by their groups and, using these as springboards, develop proposals which might illustrate the application of research methods to specific problems. From discussions so directed, it was hoped, would emerge examples of the first part of the research process specifically applied. Beyond the personal and immediate, or far-reaching but often intangible, benefits to be derived from group action, there might be recorded, especially for the benefit of beginners in research who were not present at the conference, examples of the process through which researchable questions are selected from the maze of attitudes, assumptions, unverifiable opinion, emotional reactions, controversial issues, recognized but ill-defined problems that surround the teaching of so complex a subject as "English." The process might be extended, once the questions are defined, to methods of obtaining data worth analyzing, to treatments of data which will furnish insight, and to considerations of evaluating and using the findings obtained.

Leaders and participants acted upon these suggestions in various ways. Prior to the conference, all participants received four monographs which were to contribute background related to the purposes suggested: *Research in Written Composition* by Richard Braddock and others,[1] ***Research Methods in the Language Arts*** by Carlton M. Singleton,[2] and reports of the Allerton Park and New York University Conferences on Research in the Teaching of English.[3] In addition, several groups received from their leaders preparatory assignments. For example, the leader of the group on research in teaching literature distributed a summary of selected investigations, journal articles, and related references. The group studying research in teaching reading received advance copies of working papers from the Interdisciplinary Seminar in Linguistics and Reading.[4] Because of their previous experience and these preparations, every member of the conference had some concept of its purposes and procedures. It was also true, however, that the nature of the conference demanded diverse interpretations of the stated purposes.

Such ambitious plans for a three-day conference, bringing together seventy-five persons of diverse backgrounds and interests, were in danger of self-defeat, except that everyone realized that the limitations of time and human endurance would impose their own realities upon the outcomes. An incomplete estimate of the success of the study sessions, as measured by the goals outlined above (as well as stated in the introduction), can be gleaned from the recorders' reports. Illustrative designs were achieved by the groups concerned with verbal learning and communication and with literature. The group devoted to research in the teaching of reading elected to pursue the major purposes of the conference through a less structured discussion of research methods and the findings of linguistics as they relate to the teaching of beginning reading. The groups concerned with research in teaching language and in written composition chose to explore prior questions of theories of learning and methods of teaching, discussing theoretical frameworks which are basic to consideration of specific research studies in these areas.

---

[1] Richard Braddock, Richard Lloyd-Jones, and Lowell Schoer, **Research in Written Composition** (Champaign, Illinois: National Council of Teachers of English, 1963).

[2] Carlton M. Singleton (ed.), **Research Methods in the Language Arts.** A publication of the National Conference on Research in English (Champaign, Illinois: National Council of Teachers of English, 1961).

[3] See footnote 2, page 12, and footnote 3, page 85.

[4] See page 133.

What Understandings Were Reached?

The recorders' reports are incomplete estimates of the success of the study sessions because they can represent only tangible outcomes. And even these cannot be explicitly described in limited space. Within the obvious limitations of the structure of the conference, it appears that the study groups achieved the purpose of testing (though not proving, of course) the proposition that specialists in English and in educational research can communicate on problems of mutual concern. They can reach modest goals of mutual understanding and respect. On the very limited evidence provided by the conference, it seems safe to predict that they will learn how to work productively together, though perhaps every participant shared the sober doubts expressed by Lewis Leary as he contemplated the semantic walls to be scaled.

Perhaps a more objective evaluation may be reached by summarizing ideas which converged from several different discussions to form points of agreement or common concern. Only as oversimplified generalizations, however, can these be presented as the consensus of individuals who are distinguished for the precision and independence of their thinking. Nevertheless, these trends in thinking about research were observed in various forms:

- Only research related to some kind of overarching construct, or conceptualization, of the process of education in English is likely to be of more than limited significance. See, for example, Russell's comment regarding theory, Meckel's paper on the importance of testing learning sequences, Rosenblatt's report of the New York University Seminar, which distinguishes between basic and applied studies, the statement of Steinberg and Forehand, and the recorders' reports from the groups studying research in teaching language and written composition.
- Concern for the structure of the subject matter of English and the sequential process of teaching and learning is found also in Gage's proposal for an overall paradigm of educational research

which would encompass models that are primarily substantive or methodological or logistical. Although many of the discoveries in English are the result of the work of an individual, the simple, neat, unrelated study often contributes only to the skills of the investigator and may not yield him much more than exercise. Overall planning and the coordinated efforts of many researchers are needed.

- Research in learning appears to be moving out of the laboratory and into the classroom, away from simple stimulus-response studies toward the investigation of complex, interacting relationships in the teaching-learning process. The move is hazardous. Human learning, as Meckel pointed out, takes time. Two-group, test-teach-test models are often so oversimplified that important variables go unobserved and even undefined. Mitzel offered practical suggestions for non-evaluative classroom observations. Gage warned against "black box" experiments. Harris reiterated the need for frequent data taking and continuous close observation of the study in progress. Diederich illustrated a similar point with respect to evaluation of growth in written composition. From the field of linguistic research, Marckwardt illustrated the importance of defining and protecting the sample so that findings can be accurately interpreted.

- Although seemingly contradictory to the foregoing, the plea for microstudies is really supportive of these ideas. Simplicity of design is desirable when it illuminates small truths and does not obscure others in the process. The small study can be important if seen in context as a part of a larger plan. Durrell urged the study of the "anatomy and physiology" of learning—the analysis of subskills in reading, for example. The value of well-designed microstudies is enhanced by replication.

- The values of large-scale enterprises and of replication in studies employing standard research techniques and standard measures, with uniformly rigorous control of variables and close observation and recording of pertinent phenomena, can be seen in the movement toward cooperative research exemplified in the proposed study of first grade reading. See the report of the group studying research in teaching reading.

- Similarly, the need for enriching sources of ideas, the wellsprings of assumptions to be tested, was recognized by many speakers and participants. It was highlighted in Ianni's account of the demonstration and curriculum centers now being developed with the support of the United States Office of Education.

- The logistics of research, in Gage's terminology, was the concern of many who viewed the problem from various angles. If research is to be cooperative, related to larger constructs, developed out of carefully defined systems of theory and methodology, based upon profound understanding of the nature of literature and language, for example—how is knowledge to be collated, synthesized, and made available to research workers? How are the findings of new research studies to be interpreted and disseminated? How can technological advances in other disciplines be employed in research in English education? What are the uses of computers and retrieval centers? Who tells the machines what to do? Who evaluates what they have done? The most pressing need in educational research is brainpower. Where are research workers coming from? Given limited human and financial resources, what are the priorities? Recruitment and education of potential research scholars is easier said than done. Indeed, one effect of conferences such as this is the growing realization that in educational research, problems are plentiful and ideas are cheap, but their solutions are scarce and enormously expensive.

Where Do We Go from Here?

Whether or not the San Francisco Conference was instructional as well as stimulating and inspirational is a judgment to be made by each participant. A generalized evaluation would have little validity and purpose.

Brief conferences, seminars, and "think sessions" have generative rather than sustaining powers. Their usefulness lasts as long as the sparks of enthusiasm which they ignite. But now that enthusiasm has been rekindled and ideas refined and redeveloped, this fourth conference in a series planned for the profession of English teaching will prove of little lasting value if it is followed *only* by more conferences, however improved in concept and execution. Scores of specific suggestions in this report, such as Peel's use of factor analysis in studying responses to literature or Diederich's list of some of the perils of parallel-group research, need to be tested in action.

The cause of research in the teaching of English will be moved ahead only if this conference, and others like it, result in sustained efforts on the part of the participants as individuals and as members of a profession. As illustrations of "sustained efforts" we offer the following:

1. Participants in the conference should continue on their own campuses the dialogues begun here between literary scholars, for example, and design specialists. On appropriate occasions they should ask for and use the services of research advisory panels.

2. If specialization in education is necessary, then the intensive cultivation of narrow fields must be balanced by the willingness of specialists to talk across the fences. This means preservation of at least a common core of language—the necessity for defining terms accurately and using them precisely.
3. The development of taxonomies of objectives and procedures in specific aspects of English such as written composition, literature, grammar, and language study is in itself a research enterprise. Such knowledge is basic to studies of structure, sequence, and evaluation of learning English.
4. Education of research workers in English and the continuing education of those who direct doctoral studies are long-term projects. Three-day conferences are too brief, and series of conferences suffer lapses in continuity. Needed now are such full-scale plans for organized educational efforts as post-doctoral seminars for periods of several weeks or more; fellowships that take advantage of the specialized research training available at different universities; internships for on-the-job training in practical research experiments; and one or more centers for coordinated efforts in solving problems of English learning and instruction.

The San Francisco Conference on Research Design thus focused attention not only on techniques and methodologies of research but on the underlying constructs and points of view on which significant research must be based. In raising important questions about the relationship of research to philosophies of teaching and learning, the conferees identified a truly basic issue which the profession must consider during the next few years.